**Pregunta de investigación y objetivos**

La pregunta de investigación fundamental de este proyecto se centra en el desarrollo de un sistema de detección de fraudes en transacciones de tarjetas de crédito mediante el uso de técnicas de Machine Learning. Específicamente, nos preguntamos si es posible identificar transacciones fraudulentas de manera precisa y oportuna utilizando un enfoque de aprendizaje automático y cómo podemos lograrlo de la manera más eficiente posible.

**Objetivo Principal**

El objetivo principal de este proyecto es diseñar, desarrollar e implementar un código de detección de fraudes en transacciones de tarjetas de crédito utilizando machine learning y la librería de Scikit-Learn para que se pueda identificar de manera efectiva y eficiente las transacciones fraudulentas en el "Credit Card Fraud Detection Dataset 2023" de Kaggle. Este sistema debe ser capaz de proporcionar resultados precisos y confiables en tiempo real, lo que ayudará a las instituciones financieras a tomar medidas proactivas para mitigar el riesgo de fraude.

**Objetivos Específicos**

**Preprocesamiento de Datos**: Realizar una limpieza y preprocesamiento de los datos del conjunto de datos "Credit Card Fraud Detection Dataset 2023" para garantizar la calidad y la integridad de los datos de entrada.

**Selección de Algoritmos**: Evaluar y seleccionar los algoritmos de Machine Learning más adecuados para el problema de detección de fraudes en transacciones de tarjetas de crédito.

**Entrenamiento del Modelo**: Entrenar y optimizar el modelo de Machine Learning utilizando varias técnicas como ajuste de hiperparámetros y validación cruzada, como visto en clase.

**Validación del Modelo**: Validar el modelo utilizando métricas de evaluación apropiadas, como precisión, recall, F1-score y matriz de confusión, para asegurarnos de que sea capaz de identificar transacciones fraudulentas con alta precisión.

**Datos para desarrollar el proyecto y análisis previo**

En el proyecto de detección de fraude de tarjetas de crédito, los datos son fundamentales para el éxito. Los datos deben ser representativos de la población real de transacciones, y deben tener suficientes variables para capturar la información relevante para la detección de fraude.

**Datos disponibles**

El conjunto de datos utilizado en este proyecto es un conjunto de datos de transacciones de tarjetas de crédito de 2023. El conjunto de datos contiene 568,630 transacciones, de las cuales 284,315 son fraudulentas y 284,315 son legítimas.

**Variables**

El conjunto de datos contiene 30 variables, que se dividen en dos categorías:

* Variables independientes: Estas variables describen las características de las transacciones. Incluyen la hora, la fecha, el monto de la transacción, el tipo de tarjeta, el lugar de la transacción, etc.
  + Son: id, V1-V28, Amount, Class.
  + Estas variables están escaladas, tienen una desviación estándar de 1 y una media de 0. ( id, amount y Class) no están escaladas.
  + Tienen nombres en clave del V1-V28 por temas de confidencialidad.
* Variable dependiente: Class, esta variable indica si la transacción es fraudulenta o legítima.

**Análisis previo**

Un análisis previo de los datos es importante para comprender la distribución de los datos y las relaciones entre las variables.

**Distribución de los datos**

El conjunto de datos está equilibrado a la perfección con exactamente 50% de los datos siendo transacciones fraudulentas. También se tienen más de 550 mil datos en total por lo que se puede facilitar el entrenamiento de un modelo de aprendizaje automático que pueda detectar fraudes de manera precisa. El único sesgo se podría encontrar en la cantidad de la transacción ya que la mayoría son de montos bajos.

**Relaciones entre las variables**

Se realizó un análisis de correlación entre las variables independientes y la variable dependiente. Los resultados mostraron que algunas variables, como el monto de la transacción y el lugar de la transacción, están correlacionadas con la variable dependiente.

**Forma de Evaluación**

Este modelo se está evaluando con la fórmula:

Confusion Matrix: A confusion matrix is a table that provides a detailed breakdown of the model's predictions. It includes metrics such as True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These metrics are essential for understanding the model's performance.

Accuracy: Accuracy is a basic metric that calculates the ratio of correctly predicted instances (both fraudulent and non-fraudulent) to the total number of instances. While it's a good starting point, it might not be the best metric when dealing with imbalanced datasets (which is often the case in fraud detection), where most transactions are non-fraudulent.

Precision and Recall: Precision (Positive Predictive Value) measures the percentage of true positive predictions out of all positive predictions. It helps to answer the question: "Of all the transactions predicted as fraudulent, how many are truly fraudulent?"

Recall (Sensitivity or True Positive Rate) measures the percentage of true positive predictions out of all actual fraudulent transactions. It helps to answer the question: "Of all the truly fraudulent transactions, how many did the model correctly predict?"

F1-Score: The F1-Score is the harmonic mean of precision and recall. It balances the trade-off between precision and recall, which can be especially useful when you have imbalanced data.

AUC-ROC Curve: The Receiver Operating Characteristic (ROC) curve is a graphical representation of the model's performance across different thresholds. The Area Under the ROC Curve (AUC-ROC) summarizes the model's ability to distinguish between fraudulent and non-fraudulent transactions. A higher AUC-ROC indicates better model performance.

Cross-Validation: Use techniques like k-fold cross-validation to assess the model's generalization performance. Cross-validation helps ensure that the model performs consistently on different subsets of the data.

**Conclusiones**

Los datos disponibles para el proyecto son prometedores ya que contienen un 50% de datos fraudulentos. Sin embargo, se debe revisar bien el conjunto de datos no tenga valores duplicados o por el estilo, que puedan dificultar el entrenamiento de un modelo de aprendizaje automático preciso. Se necesita más análisis para comprender mejor las relaciones entre las variables.

**Detalles para considerar**

El conjunto de datos se recopiló durante un período de tiempo determinado. Es importante tener en cuenta que las tendencias de fraude pueden cambiar con el tiempo. Por lo que la efectividad del modelo puede que no se mantenga con el tiempo. También es importante notar que hay diferentes tipos de fraudes, uno puede ser el que el banco rechaza por ser fraudulento, el otro puede ser cuando se utiliza la tarjeta sin el consentimiento del usuario y el cobro se realiza. Si se tuviera dos categorías de fraudes de pronto la precisión del modelo podría mejorar.