**Aprendizaje de Máquina Aplicado**

**Fecha: 5 de octubre de 2023**

**Universidad EAFIT**

**Medellín, Colombia**

**Detección de Fraudes en Tarjetas de Crédito**

**Estudiantes:**

**Sebastian Arteaga**

**Juan Felipe Salamanca**

**Efraín Hernández**

**Pregunta de investigación y objetivos**

La pregunta de investigación de este proyecto se centra si es posible identificar transacciones fraudulentas de manera precisa, correcta y oportuna utilizando un enfoque de aprendizaje automático, y que métodos nos permiten 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" encontrado en 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**

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 primero varias técnicas de regresión logística, después intentaremos modelos como; árbol de decisiones, redes neuronales y modelos de aumento de gradiente, entre otros.

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.

Minimizar falsos negativos: Minimizar la cantidad de transacciones que se dice que no fueron fraude, pero si lo fueron.

**Metodología de Investigación**

La metodología de investigación propuesta se centra en la aplicación y comparación de diversos algoritmos de aprendizaje automático con el objetivo de identificar cuál de ellos ofrece los mejores resultados en términos de precisión y eficiencia. Los algoritmos seleccionados para esta investigación son: Random Forest, Regresión Logística, Regresión Logística Polinomial, Gradient Boosting, MLP Classifier y Gaussian NB.

Para garantizar una evaluación exhaustiva y objetiva de cada algoritmo, se implementará el método de validación cruzada mediante "Grid Search CV". Esta técnica permite explorar múltiples combinaciones de parámetros para cada algoritmo y determinar cuál es la configuración óptima que maximiza su rendimiento.

Evaluación Inicial: En esta fase, se aplicarán todas las técnicas mencionadas anteriormente para obtener una visión preliminar de su rendimiento. Esta evaluación nos permitirá identificar cuáles son los algoritmos más prometedores en términos de precisión y velocidad.

Iteraciones Adicionales: Basándonos en los resultados de la fase inicial, se seleccionarán los algoritmos que hayan demostrado ser más eficientes y precisos. Estos algoritmos serán sometidos a al menos 10 iteraciones adicionales para afinar su rendimiento y obtener resultados más robustos.

El objetivo final de esta metodología es no solo identificar el algoritmo que ofrece los mejores resultados, sino también comprender las ventajas y limitaciones de cada técnica, permitiendo así una toma de decisiones informada sobre qué algoritmo es el más adecuado para aplicaciones futuras.

**Revisión de la Literatura, Estado del Arte**

En la búsqueda de soluciones para el problema, se encontraron múltiples investigaciones que han logrado resultados cercanos a la perfección. Dada la amplia gama de métodos disponibles en la biblioteca sklearn para la clasificación, se realizó una revisión exhaustiva para comprender las características y aplicaciones de cada uno.

Random Forest: Es un método de ensamblaje que combina múltiples árboles de decisión para mejorar la precisión y evitar el sobreajuste. Es especialmente efectivo en conjuntos de datos grandes y con muchas características.

Regresión Logística y Regresión Logística Polinomial: Son técnicas de clasificación binaria que estiman la probabilidad de que una instancia pertenezca a una clase particular. La versión polinomial permite capturar relaciones no lineales en los datos.

Gradient Boosting: Es un algoritmo de ensamblaje que construye árboles de decisión de manera secuencial, donde cada árbol corrige los errores del anterior.

MLP Classifier: Es una red neuronal de perceptrón multicapa que puede capturar relaciones complejas en los datos.

Gaussian NB: Es un clasificador basado en el teorema de Bayes que asume que las características siguen una distribución normal.

En cuanto a las métricas de evaluación en sklearn, se destacan:

Accuracy: Mide la proporción de predicciones correctas.

Precision: Indica la proporción de identificaciones positivas que fueron realmente correctas.

Recall: Muestra la proporción de positivos reales que fueron identificados correctamente.

F1-Score: Es una medida que combina precisión y recall.

Confusion Matrix: Proporciona una visión detallada del rendimiento del clasificador.

ROC AUC: Evalúa la capacidad del modelo para discriminar entre clases.

Cross Validation: Es una técnica que divide el conjunto de datos en múltiples subconjuntos para entrenar y probar el modelo, asegurando una evaluación más robusta.

Una investigación relevante en este ámbito es la realizada por S. Bhat, donde se aplicaron diversos algoritmos de aprendizaje, incluyendo Gradient Boosting, Random Forest y Regresión Logística, para la detección del Síndrome de Ovario Poliquístico. En este estudio, se evaluaron los modelos utilizando métricas como Accuracy, Precision, Recall, F1-score, ROC curve plot, AUC score y K Fold Cross validation1. Aunque este era un tema distinto pudimos ver que tenían similitudes en la utilización de modelos y métricas de evaluación.

**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. Los datos también deben estar bien balanceados para no tener sesgos, es decir no debemos tener muchos mas datos de no fraudes a datos de fraudes.

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. Algunas de estas contienen valores repetidos en algunas columnas por lo que seria bueno considerar una depuración de los datos que tengan valores repetidos.

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. Pero están guardadas con nombres claves para proteger los datos.

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.

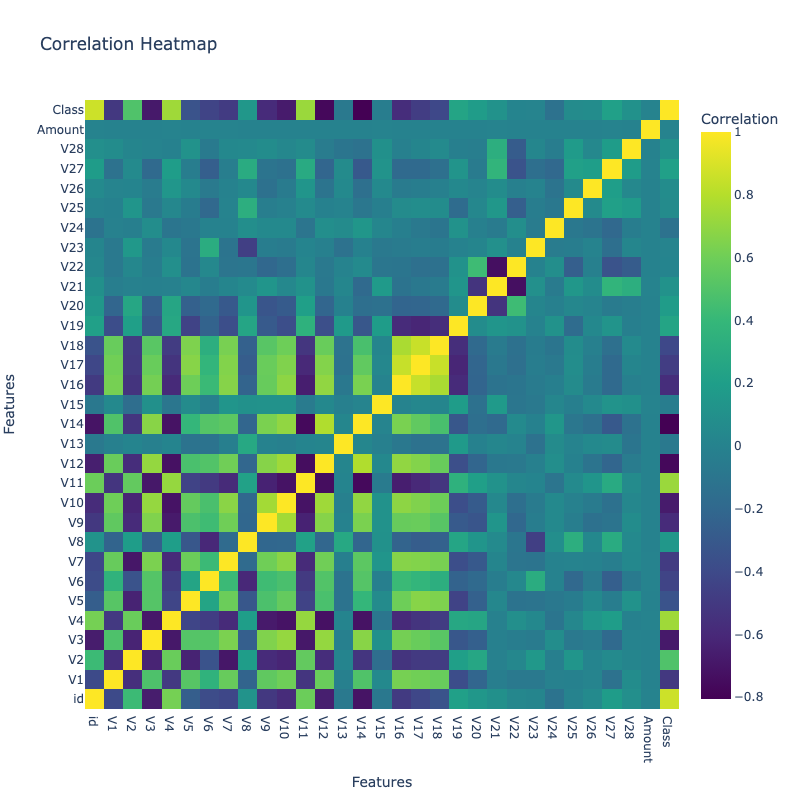
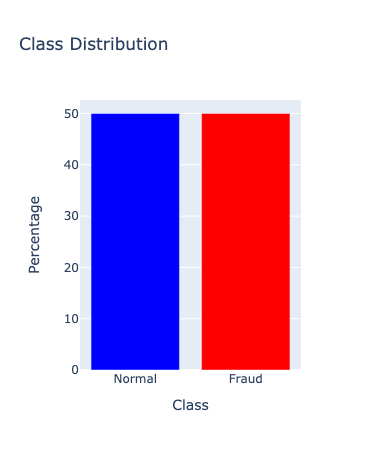
Amount no esta escalada.

Tienen nombres en clave del V1-V28 por temas de confidencialidad.

Variable dependiente es Class, esta variable indica si la transacción es fraudulenta o legítima. (0 no es fraude, 1 es fraude)

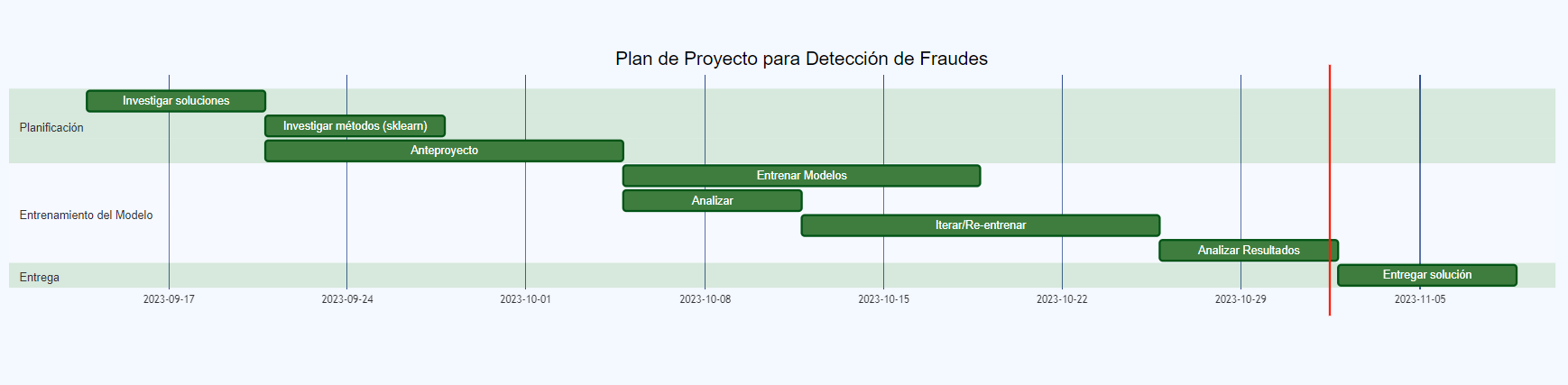
El presente data set no posee variables muy diversas, y la gran mayoría están ocultadas por confidencialidad. La única variable a la cual se llegó a la conclusión que debía tener un análisis previo es la de Class la cual se coloca en 1 cuando es fraudulenta y cuando no es 0 se puede ver la en la figura más abajo.

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. Los montos de las transacciones también están bien distribuidos entre valores de 50 a 24.000 USD, por lo que se tiene una fuente de datos muy prometedora.

Se realizó un análisis de correlación entre las variables independientes y la variable dependiente. Los resultados mostraron que algunas variables, tienen correlaciones altas tanto positivas como negativas. Es importante notar que la que obtuvo la correlación más alta es ID, y esta se debe ignorar ya que esta correlación debe ser pura coincidencia de la forma en la que se creó la base de datos. Se puede observar el grafico de las correlaciones abajo:  

Los datos disponibles para el proyecto son prometedores ya que contienen un 50% de datos fraudulentos, no tienen sesgos y muchas de las columnas tienen buenos valores en las correlaciones. 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.

**Plan**



**Implicaciones éticas**

La implementación de algoritmos de aprendizaje automático en la detección de transacciones fraudulentas, aunque prometedora, conlleva importantes consideraciones éticas. Al utilizar datos encontrados en Kaggle, se garantiza cierta confidencialidad, ya que las variables han sido anonimizadas o transformadas para proteger la identidad y la información sensible de los individuos. Sin embargo, es esencial considerar las siguientes implicaciones:

Si el algoritmo demuestra ser eficiente en la detección de transacciones fraudulentas, podría reducir significativamente la carga de trabajo manual en esta área. Aunque esto puede ser visto como una ventaja, también podría llevar a la reducción de empleos o al cambio de roles dentro de las organizaciones. Es esencial que las transacciones detectadas como fraudulentas sean revisadas por expertos humanos. Los algoritmos, por avanzados que sean, pueden cometer errores, y es crucial garantizar que no se tomen medidas drásticas basadas únicamente en predicciones automáticas. Con el tiempo, los modelos de aprendizaje automático pueden perder precisión debido a la evolución y cambios en los patrones de transacciones. Esto requiere una revisión y reentrenamiento periódico del modelo, lo que implica una responsabilidad continua por parte de las organizaciones que lo implementen.

**Aspectos Legales y Comerciales**

Desde una perspectiva comercial, la implementación exitosa de este algoritmo tiene un potencial significativo, tanto dentro como fuera de la compañía. La detección automática y precisa de transacciones fraudulentas puede resultar en ahorros considerables y en una mayor confianza por parte de los clientes.

Sin embargo, es esencial considerar los aspectos legales relacionados con el uso de datos y la toma de decisiones basadas en predicciones automáticas. Aunque los datos de Kaggle están bien presentados y anonimizados, es crucial garantizar que cualquier implementación cumpla con las regulaciones de protección de datos y privacidad. Además, las decisiones tomadas a partir de las predicciones del algoritmo deben ser transparentes y justificables, especialmente si tienen implicaciones legales o financieras para los individuos afectados.

**Referencias**

Bhat, S. (2021). Detection of Polycystic Ovary Syndrome using Machine Learning Algorithms. Recuperado el 16 de agosto de 2021, de https://norma.ncirl.ie/5137/1/shakoorahmadbhat.pdf.