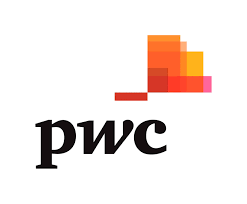
INDICE DEL INFORME PROPUESTA SOLUCION RETO PwC



[Predicción del Fraude en Tarjetas de Crédito – Reto PwC 3](#_Toc151564592)

[Introducción 3](#_Toc151564593)

[Objetivo 3](#_Toc151564594)

[Pasos para Predecir Fraude en Tarjetas de Crédito 3](#_Toc151564595)

[A. Adquisición de Datos 3](#_Toc151564596)

[B. Análisis Exploratorio de Datos (EDA) 5](#_Toc151564597)

[C. Ingeniería de Características 5](#_Toc151564598)

[D. Preparación de Datos para Modelado 5](#_Toc151564599)

[E. Selección y Entrenamiento del Modelo 5](#_Toc151564600)

[F. Evaluación del Modelo 5](#_Toc151564601)

[G. Validación y Optimización del Modelo 5](#_Toc151564602)

[H. Despliegue y Monitoreo 5](#_Toc151564603)

[Conclusiones 6](#_Toc151564604)

[Consideraciones Finales 6](#_Toc151564605)

[A. Objetivos de la actividad : 7](#_Toc151564606)

[B. Enunciado de la Práctica, se pide: 7](#_Toc151564607)

[Estructura proyecto 8](#_Toc151564608)

[Task00 8](#_Toc151564609)

[Task01 8](#_Toc151564610)

[Task02 9](#_Toc151564611)

[Task03 11](#_Toc151564612)

[Task04 12](#_Toc151564613)

[Task05 12](#_Toc151564614)

[a) Qué deberá entregar / subir al CANVAS ¿? : 12](#_Toc151564615)

[b) Nombrado de archivos e indicaciones de cómo subir y formato: 12](#_Toc151564616)

# Predicción del Fraude en Tarjetas de Crédito – Reto PwC

## Introducción

La detección de fraude en transacciones con tarjetas de crédito es un desafío constante para las instituciones financieras. El uso de técnicas de machine learning puede ser fundamental para identificar patrones y anomalías en los datos que podrían indicar actividades fraudulentas.

### Objetivo

Este informe describe los pasos esenciales, que realizaremos sobre el dataset (ver punto A ) para predecir el fraude en transacciones con tarjetas de crédito utilizando técnicas de machine learning con diferentes herramientas, las mismas que se justifican por la premura en la entrega de presente informe, ;entre otras:

1. RapidMiner
2. BigQuery
3. Cloud

## Pasos para Predecir Fraude en Tarjetas de Crédito

## Adquisición de Datos

Obtención de datos:

Nos entregan un dataset con transacciones de tarjetas de crédito, no hay fechas en las transacciones, incluyen características como cantidad, tipo de transacción, entre otros.

#### Descripción del dataset original del reto PwC

1. Nombre: full\_dataset.csv
2. Tipo: CSV
3. Observaciones : 6.362.620 observaciones
4. Atributos: 11 atributos.

* A screenshot of a computer

  Description automatically generated

Limpieza y preprocesamiento: Eliminar datos duplicados, manejar valores faltantes y transformar variables si es necesario (normalización, codificación de variables categóricas.

### Eliminar datos duplicados

No hay duplicados.

### Valores faltantes

No hay duplicados valores faltantes

## Análisis Exploratorio de Datos (EDA)

Exploración de características: Analizar la distribución de variables, identificar correlaciones y buscar posibles patrones o anomalías en los datos.

Visualización de datos: Utilizar gráficos y visualizaciones para entender mejor la distribución de transacciones normales y fraudulentas.

### ¿En que categorías está el fraude?

A screenshot of a computer

Description automatically generated

### Dónde están los atributos que no aportan nada?

Es ínfimo la cantidad de observaciones, por lo tanto, eliminamos este atributo.

A screenshot of a cell phone

Description automatically generated

## Ingeniería de Características

Selección de características: Identificar las características más relevantes para el modelo predictivo.

Creación de nuevas características: Derivar características adicionales que puedan mejorar la capacidad predictiva del modelo.

## Preparación de Datos para Modelado

División de datos: Separar los datos en conjuntos de entrenamiento, validación y prueba.

Balanceo de datos: Si es necesario, aplicar técnicas de remuestreo para equilibrar clases (oversampling, undersampling, SMOTE, etc.).

En la clase anterior, realizamos el under sampling, el método de over sampling lo dejamos para más adelante, dado que requerimos gran cantidad de recursos y lo llevaremos a la nube.

## Selección y Entrenamiento del Modelo

Elección del algoritmo: Seleccionar modelos de machine learning adecuados para la predicción de fraude (Random Forest, Support Vector Machines, Redes Neuronales, etc.).

Entrenamiento del modelo: Utilizar el conjunto de entrenamiento para entrenar el modelo seleccionado.

## Evaluación del Modelo

Validación del modelo: Evaluar el rendimiento del modelo utilizando métricas como precisión, exhaustividad, F1-score, matriz de confusión, ROC-AUC, entre otras.

Ajuste de hiperparámetros: Optimizar los hiperparámetros del modelo para mejorar su desempeño.

## Validación y Optimización del Modelo

Validación cruzada: Verificar la generalización del modelo utilizando técnicas de validación cruzada.

Optimización adicional: Realizar ajustes adicionales en el modelo para mejorar su capacidad predictiva.

## Despliegue y Monitoreo

Implementación del modelo: Integrar el modelo en un entorno de producción para monitorear transacciones en tiempo real, ya que en la actualidad solo podrá predecir entorno simulados o contratos, no en tiempo real.

Monitoreo continuo: Supervisar el rendimiento del modelo y realizar actualizaciones según sea necesario para mantener su efectividad.

### Demo funcional

[Check4U](http://pwc-chalenge.firebaseapp.com/)

Repositorio del código

https://github.com/sukuzhanay/m3.git

## Conclusiones

La predicción de fraude en tarjetas de crédito mediante técnicas de machine learning es fundamental para mitigar riesgos financieros. Los pasos mencionados constituyen un marco sólido para desarrollar un sistema efectivo de detección de fraudes.

### Consideraciones Finales

La actualización constante del modelo es crucial para adaptarse a nuevos patrones de fraude.

La colaboración con expertos en seguridad financiera es esencial para mejorar la precisión y eficacia del modelo.

PRACTICA RETO PWC

En esta práctica los alumnos deberán explorar el Datasets suministrado por el profesor y que se encuentra en el Canvas, se apoyaran en la la plataforma de data science RapidMiner y conforme a las practicas realizadas en clase, deberán poner en práctica lo visto para generar un documento Word que responda a las preguntas de negocio que figuran en el enunciado.

El que mejor predicción obtenga, será el que mejor nota obtenga.

# Objetivos de la actividad :

1. Conocer como se realiza un ETL + EDA + Modeling
2. Aplicar la plataforma RapidMiner para resolver preguntas de negocio.
3. Entender como funcionan los modelos y elegir el de mejor rendimiento
4. Generar el documento que deberá ser subido al Canvas

# Enunciado de la Práctica, se pide:

**Caso de uso:**

Basados en lo aprendido hasta la fecha y en las prácticas realizadas en clase deberá responder a las preguntas que cumplan los siguientes requerimientos:

## Estructura proyecto

## Task00

Para ello, primero vamos a definir la estructura de nuestro repositorio y crearemos los siguientes folders.

A white background with black text

Description automatically generated

Donde :

**ue21535220**: es el folder del proyecto, cada estudiante creara el suyo con el fomato: uenumexp ejemplo( ue21525220).

**Data**: se guardarán todos los data sets

**Models**: almacenarán los modelos

**Process**: almacenarán todos los procesos

**Reports**: guardaremos los informes, imágenes, PDFs

Results: guardaremos los resultados resultantes de la aplicación de los procesos

## Task01

Guarde y lea el archivo el archivo **full\_dataset.csv** y guárde el archivo resultante de la lectura como exampleset en un repositorio llamado uenumexp ejemplo( ue21525220), dentro de la carpeta Data, con el nombre whole\_dataset, para su posterior utilización.

A screenshot of a computer

Description automatically generated

## Task02

Como podrá observar el data set está completamente desbalanceado en relación a la variable **isFraud.**

**A screenshot of a graph

Description automatically generated**

**Balancéalo**, todo el dataset en relación a la variable **isFraud** sin perder INFORMACION en la clase minoritaria, para que en la etapa de modelado no tengamos problemas de over y/o under fitting.

Para ello, utilice los operadores adecuados; deberá obtener un nuevo **exampleset** que deberá guardarlo dentro del folder **Data**, el nombre: **balanced\_dataset**

A screenshot of a computer

Description automatically generated

Recuerde, debe estar balanceado como se puede apreciar en la figura siguiente.

## Task03



## Task04

Pruebe al menos **tres modelos** y elija el de mejor rendimiento.

## Task05

Realice los pasos del 1-3, usando Python, tal y como lo hemos hecho en clase, adjunte el archivo resultante con el nombre **ue21535220.ipynb**

# Qué deberá entregar / subir al CANVAS ¿? :

* 1. Memoria descriptiva PDF, con portada, índice de la actividad, donde figuren las respuestas a las preguntas del enunciado, así como todos los gráficos / mapas / informes generados, capturas. (debe constar el nombre del alumno, con su enlace al repositorio donde esta el código entregado, IMPORTANTE el enlace debe estar activado (hyperlink).
  2. Se deberá explicar los procesos y decisiones tomadas y el porque, caso contrario, la actividad se considera no entregada.
  3. Todo el repositorio deberá subirse al repositorio de GitHub de cada alumno y añadirme como colaborador ( con el rol de propietario para poder evaluar) al repositorio de vuestro GitHub; mi username es : [**sukuzhanay@gmail.com**](mailto:sukuzhanay@gmail.com) ( **si no me añadís, se considera no entregado**)
  4. La entrega es individual
  5. **IMPRESCINDIBLE** : los archivos deberán ser subidos a vuestros repositorios .
  6. Fecha de entrega: según figure en el Canvas.

# Nombrado de archivos e indicaciones de cómo subir y formato:

* 1. Todos los archivos entregados deberán subirse de forma individual, NO comprimidos (zip, tar, etc)