RAPIDMINER

An ALTAIR Company

Práctica ETL + EDA + MODELING

A. Introducción

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.

B. 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

C. Enunciado de la Práctica, se pide:

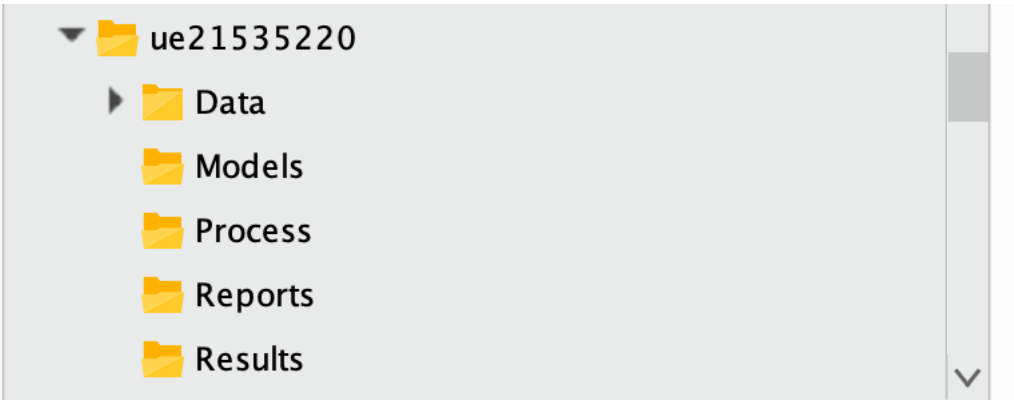
Caso de uso:

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.

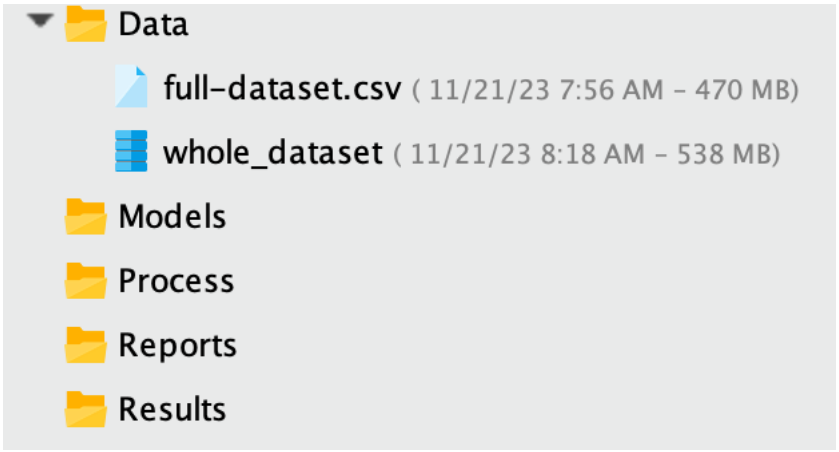


Donde :

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

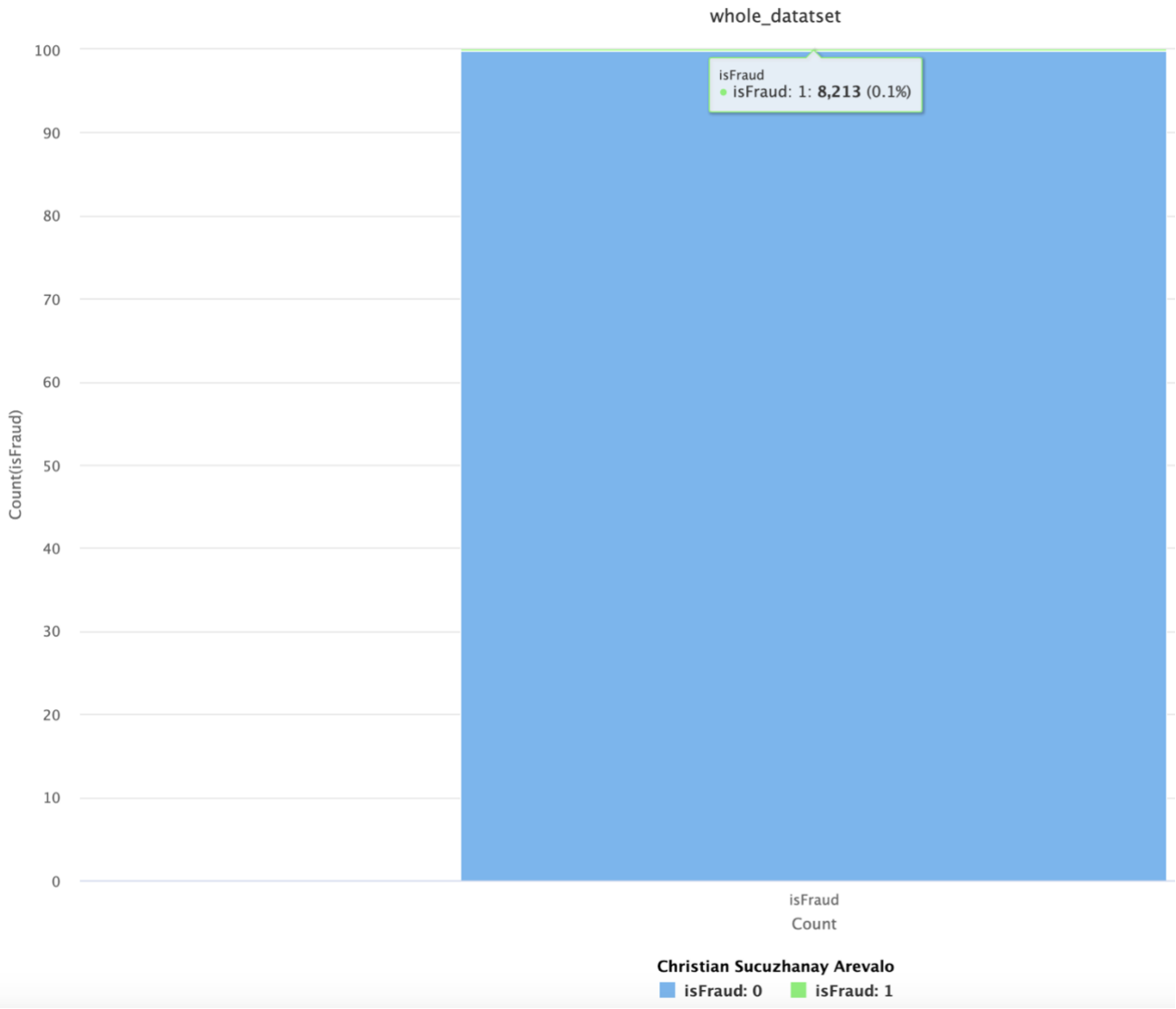
Task01

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.



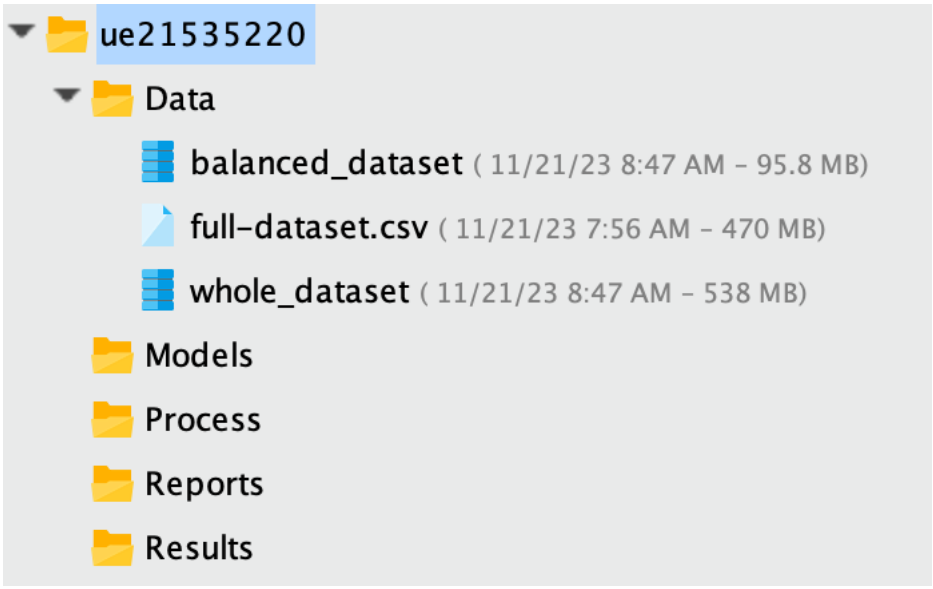
Task02

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



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**



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**

a) Qué deberá entregar / subir al CANVAS ¿? :

a. 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).
b. Se deberá explicar los procesos y decisiones tomadas y el porque, caso contrario, la actividad se considera no entregada.
c. 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** (**si no me añadís, se considera no entregado**)
d. La entrega es individual
e. **IMPRESCINDIBLE** : los archivos deberán ser subidos a vuestros repositorios .
f. Fecha de entrega: según figure en el Canvas.

b) Nombrado de archivos e indicaciones de cómo subir y formato:

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

Christian Vladimir Sucuzhanay Arévalo