DIPLOMADO Ciencia de Datos

Examen

Nombre: Rodríguez (Apellido Paterno) Fitta (Apellido Materno) José Emanuel (Nombre(s)) **Fecha**: 26/02/2022

Calificación: _____

Objetivo: *El participante identificará los objetivos y fases del modelo CRISP-DM para un caso de estudio proporcionado por el participante.*

Instrucciones:

1. Llenar cada uno de los recuadros siguientes explicando el caso de estudio a tratar, así como sus objetivos y los objetivos de minería de datos.
2. Llenar cada uno de los recuadros explicando cada una de las fases del modelo CRISP-DM para el caso de estudio elegido por el participante.

Explicar el caso de estudio: **Valor 1 punto**

Una de las funciones de una institución financiera es el otorgamiento de tarjetas de crédito a sus clientes, sin embargo, al hacerlo ha detectado que cierto porcentaje de estos han hecho fraude con ellas y mediante el robo de identidad cargan sus compras a perfiles falsos o a la cuenta de clientes reales. Esto es importante, debido a que esta situación genera pérdidas económicas considerables.

Objetivos del caso de estudio: **Valor 1.5 puntos**

El objetivo del análisis es identificar clientes que puedan cometer fraude y con ello tomar medidas como pueden ser trackear el uso que se la da a su tarjeta y confirmar que en efecto están cometiendo fraude y limitar el uso de la tarjeta o en su caso incluso, bloquear dicha tarjeta. Esto beneficiaría a la empresa, puesto que se disminuiría en gran medida la cantidad de pérdidas económicas.

Objetivos de la minería de datos: **Valor 1.5 puntos**

Se busca encontrar las principales características de aquellas personas que han cometido fraude, para con ello crear un algoritmo que sea capaz de predecir a los futuros clientes que puedan cometer fraude.

Fases del modelo CRISP_DM

1. Comprensión del negocio

Valor 1 punto

Esta es una empresa dedicada al otorgamiento de tarjetas de crédito, los fraudes realizados, ascienden al 1 %, dado que esta es una empresa con gran cantidad de clientes, se estima que a pesar de ser solo un 1%, esto genera pérdidas económicas importantes.

2. Comprensión de los datos

Valor 1 punto

La base de datos que se tiene consta principalmente de los movimientos hechos por los clientes como lo son la cantidad total de dinero que se deja para compras, los pagos mínimos, los adelantos que cada cliente da, el límite de crédito, etc., los datos son de tipo numérico, además esta base no contiene datos sensibles como lo son la información personal. La tabla tiene un registro distinguiendo aquellos clientes que se ha podido corrobar que cometieron fraude en el pasado de los que no. Se tiene un registro de la fecha en que se activo la cuenta, un registro de su último pago.

3. Preparación de los datos

Valor 1 punto

En este caso se considerará la eliminación de valores nulos en la columna de las fechas y se aplicará una técnica de oversampling para balancear los datos debido a que como se mencionó se tiene tan solo un 1% de casos de fraude. En el caso de valores nulos en el resto de columnas se reemplazarán con el promedio.

4. Modelado

Valor 1 punto

Se hará un modelo predictivo, aplicando algoritmos de Machine learning como son LogisticRegrssion, Naive-Bayes, XGBoost classification y árboles de decisión.

5. Evaluación

Valor 1 punto

Para el evaluar el modelo que es el que mejor desempeño tiene se utilizará la métrica recall_score puesto que es importante que la cantidad de falsos negativos sea mínima, puesto que el no detectar personas que realmente cometen fraude podría significar en seguir teniendo pérdidas económicas

6. Despliegue y explotación

Valor 1 punto

Una vez terminado el modelo se realizará un despliegue el cual debe irse actualizando cada semana, esto debido a que los métodos de fraude cambian constantemente.

Valor total 10 puntos