

**Prueba de Analista  
Científico de Datos**

 **Banistmo**

# OBJETIVO

Medir el pensamiento crítico y creatividad analítica del participante mediante el desarrollo de un modelo estadístico computacional que logre los siguientes resultados:

1. Desarrollar un Modelo para calcular:
  - La Probabilidad de Impago (i.e. Probabilidad de Default) de un cliente [Variable Continua: 0.0 – 1.0]
  - El estatus de default [variable Binaria: Si=1, o No=0]
2. Estimar el monto total (en unidad de dinero) de la cartera en perdida.
3. Exponer el modelo en un API para consumir el modelo (Extra puntos)

# DATASET

## 1. Set de datos: *default\_dataset.csv*

- Descripción:

El set de entrenamiento presentado contiene información de clientes de algún banco en el que se mide diferentes variables del sistema y del estatus de impago, también conocido como default, del cliente.

- Variable Objetivo: default [**Binaria (Si=1, No=0)**]
- Variables Explicativas: 24 [**Mixtas (Continuas y Categóricas)**]

# CRITERIOS DE EVALUACIÓN

1. Lograr una métrica de rendimiento generalización sin sobre entrenar el modelo utilizando el dataset de evaluación el cual es oculto para el candidato.
  - Si bien existen varios tipos de métricas de clasificación, solo un subgrupo de ellas son aptas para el problema presentado en esta prueba. Escoger una y sustentar su escogencia.
  - La magnitud de la métrica de rendimiento de exactitud deberá ser igual o mayor que el modelo de referencia (oculto).
2. Documentación de la metodología de desarrollo que contemple (En caso de aplicar):
  - Metodología Limpieza de datos
  - Metodología de Selección de variables
  - Metodología de Imputación
  - Metodología de Escogencia de algoritmo de modelación
  - Metodología de Optimización de hiper parámetros
  - Metodología de evaluación
  - Metodología de estimación de la perdida esperada, en unidad de dinero, predicha por el modelo.
3. Adicionalmente, el dataset cuenta con problemáticas que surgen por la naturaleza del problema y estas dictan las decisiones de modelación. El participante debe documentar y sustentar sus decisiones.

# EXTRA PUNTOS

1. Para la obtención de extra puntos y lograr ser considerado con mayor probabilidad de éxito el candidato podrá optar por realizar un HTTP API que exponga el modelo desarrollado para su consumo con las siguientes especificaciones:

- GET : `url/model/v1/healthCheck`
- POST: `url/model/v1/predict/user_id`
- Json response example:

```
{“y_label”: 1,  
  “y_pred”: 0.85,  
  “metadata”: {“timestamp”: 19584325,  
               “version”: 1}  
}
```

- El lenguaje para el desarrollo del server aplicativo debe ser en Python.

# INSTRUCCIONES

1. Lenguaje de programación es Python
2. Método de presentación de la respuesta es un Jupyter notebook
3. Los sustentos textuales deben estar presentes como comentarios o markup dentro del Jupyter Notebook
4. El candidato tiene exactamente 10 días continuos para entregar la prueba resuelta contados a partir del envío de la misma
5. Banistmo tomará hasta 5 días hábiles en calificar la prueba
6. Una vez calificada se contactará al participante el resultado obtenido y de ser exitoso pasar a una segunda etapa de entrevista técnica por video conferencia