# TALLER N°3 MACHINE LEARNING: BANCO STELLAR

## A. Objetivo General:

- Desarrollar y aplicar técnicas avanzadas de aprendizaje supervisado en Machine Learning para resolver un desafío crítico de clasificación de clientes potenciales de alto perfil para el Banco Estelar, considerando una estructura de costos y beneficios específica. El proyecto debe enfocarse en maximizar la precisión del modelo, minimizando tanto los falsos positivos como los negativos, y así garantizar la eficiencia financiera y la efectividad de la estrategia de marketing del banco.

## B. Objetivos Específicos:

1. Modelar un problema del mundo real asociado a datos de una entidad financiera.
2. Aplicar un enfoque estructurado para abordar el modelamiento basado en CRISP-DM.
3. Modelar con 3 algoritmos de ML distintos el mismo problema y seleccionar el mejor algoritmo para la situación, basándose en métricas de desempeño.
4. Justificar cada decisión del modelamiento que conduzca a generar mejores resultados.
5. Como mínimo se espera que el equipo de estudiantes logre:
   - **Análisis Exploratorio de Datos (EDA)**:
     - Realizar un análisis exhaustivo de los datos proporcionados por el Banco Estelar, enfocándose en identificar patrones, tendencias y relaciones clave que puedan influir en la clasificación de clientes de alto perfil.
   - **Pre-procesamiento de Datos:**
     - Aplicar técnicas de limpieza y normalización de datos para garantizar la calidad y la consistencia del conjunto de datos.
     - Manejar valores faltantes, atípicos y errores en los datos de forma efectiva para mejorar la precisión del modelado.
     - Codificar adecuadamente las variables categóricas y estandarizar las variables numéricas para optimizar el rendimiento de los modelos de aprendizaje automático.
   - **Ingeniería y Selección de Características:**
     - Desarrollar características innovadoras basadas en los datos disponibles para mejorar la capacidad predictiva de los modelos.
     - Seleccionar las características más relevantes y efectivas, considerando su impacto en la precisión y la interpretabilidad del modelo.
   - **Modelado Predictivo con Tres Algoritmos Diferentes:**
     - Construir y ajustar modelos utilizando XGBoost, Light GBM y Redes Neuronales.
     - Evaluar y comparar los modelos basándose en métricas de rendimiento relevantes como precisión, recall, F1-score y análisis de la matriz de confusión, con un enfoque particular en minimizar los costos asociados con clasificaciones incorrectas.
   - **Validación y Pruebas Rigurosas:**
     - Implementar técnicas de validación cruzada para garantizar la robustez y la generalización de los modelos.
     - Realizar pruebas exhaustivas utilizando un conjunto de datos de prueba independiente para evaluar la efectividad de los modelos en condiciones reales.
   - **Análisis de Impacto Financiero:**
     - Analizar el impacto financiero de los modelos predictivos, considerando el costo de adquirir nuevos clientes y el beneficio potencial de identificar correctamente a los clientes de alto perfil.

- Desarrollar un marco para cuantificar y comparar el retorno de la inversión (ROI) de los diferentes modelos.

- **Justificación y Reflexión Ética:**
  - Justificar cada decisión tomada en el proceso de modelado, desde la selección de características hasta la elección del algoritmo, con un enfoque en la ética y la transparencia.
  - Reflexionar sobre las implicaciones éticas y los posibles sesgos en los modelos, proponiendo estrategias para mitigarlos.

## C. Contexto de NegocioBanco Stellar

**Introducción**

En un audaz movimiento para conquistar el emergente mercado financiero chileno, un prestigioso banco suizo, apodado el "Banco Stellar", se embarca en una misión que busca transformar su enfoque tradicional hacia una estrategia orientada a datos, con el objetivo principal de identificar y capturar un segmento de mercado personas exclusivo: clientes de alto perfil en Chile, aquellos cuyos ingresos netos anuales superan los 60 millones de pesos.

**La Trama**

El Banco Stellar, conocido por su innovación y liderazgo en la industria financiera, ha reunido un conjunto de datos exhaustivo, que abarca desde información demográfica hasta detalles financieros íntimos de potenciales clientes. Sin embargo, la mina de oro de datos es solo el comienzo. Aquí es donde entran en juego los ingenieros de Machine Learning, convocados de diversas consultoras para enfrentar el desafío de "Operación Data Nova".

**La Misión**

Como uno de los Equipos de Ingenieros de ML seleccionados, se te encomienda una tarea de alta envergadura: desarrollar un modelo predictivo no solo preciso y eficaz, sino también económicamente viable. Este modelo debe identificar a aquellos individuos que no solo cumplen con el umbral de ingresos, sino que también representan una oportunidad de inversión lucrativa para el banco, equilibrando astutamente los costos de adquisición (7 UF) con las ganancias potenciales (70 UF anuales).

## D. Dataset para el Modelamiento ML

El dataset "**highprofilecustomers.csv**" (HPC), representa un compendio vital de datos recopilados por el banco en su ambiciosa expansión hacia el mercado chileno. Este conjunto de datos ha sido meticulosamente curado por investigadores de mercado para identificar potenciales clientes para servicios exclusivos de inversión, enfocándose en aquellos con ingresos anuales superiores a los 60 millones de pesos chilenos.

Dada la importancia de precisión en la selección de clientes, se le ha encomendado a su equipo de ML, la misión crítica de construir el modelo más eficaz utilizando "High Profile Customers". Su habilidad y competencia como equipo serán evaluadas rigurosamente a través del desempeño de su modelo en un conjunto de datos de prueba, que permanecerá oculto hasta la fase final de evaluación.

El costo de adquisición de un nuevo inversor es de 7 UF, mientras que el retorno promedio se estima en unos 70 UF anuales por cada inversor objetivo seleccionado correctamente. Por lo tanto, la precisión y la eficacia en la identificación de estos perfiles de alto valor es de suma importancia.

El dataset High Profile Customers se compone de las siguientes características:

OBS: las variables no tienen tilde a propósito para ser usadas en Python/ Jupyter notebook.

1. **edad**: número de años enteros que tiene la persona.
2. **tipo_empresa**: indica el tipo de empresa u organización en la que trabaja la persona.
3. **peso_final**: es número de unidades en la población objetivo que representa para el banco la persona. Las unidades no tienen medida es un indicador a-dimensional.
4. **max_nivel_educ**: es el máximo nivel de educación al que atendió la persona.
5. **agnos_educacion**: es el número de años de estudios completos que realizó la persona.
6. **estado_civil**: representa en qué situación marital se encuentra la persona.
7. **ocupacion**: representa el tipo de rol o cargo que tiene la persona.
8. **relacion_jhogar**: indica que en tipo de relación matrimonial o familiar está esta persona.
9. **ciudadania**: esta característica indica si la persona.
10. **sexo**: indicación del género de la persona.
11. **ganancia_cap**: es un monto de dinero (en miles de pesos) que la persona ha logrado ganar en instrumentos de inversión.
12. **perdida_cap**: es un monto de dinero (en miles de pesos) que la persona ha perdido en instrumentos de inversión.
13. **hrs_sem**: número de horas a la semana por las cuales está contradado o trabaja
14. **comuna**: comuna donde reside la persona.
15. **net60mn**: etiqueta asignada en función de certificados de ingresos y sueldos que la persona obtiene más de 60 millones [‘>60Mn’] al año o si no lo hace [‘<=60Mn’].

**E. En que consiste el Proyecto** – Esto es lo que usted debe hacer !

Como se mencionó en los objetivos, usted debe construir el mejor modelo posible que el área de Inversiones de este banco usará de forma automatizada para ofrecer productos de inversión a aquellas personas que el algoritmo ML que usted construya prediga ingresos anuales por 60 millones de pesos o más.

Para lo anterior se pide (como mínimo):

a) Usando Python con jupyter notebook construya sus análisis de los datos y el modelo que propondrá.
b) Debe realizar un Análisis Exploratorio detallado de los Datos (EDA) que permita identificar relaciones entre las variables, qué variables podrían contribuir al objetivo buscado, qué variables no aportarían, qué relaciones tienen entre ellas las variables, que variable que no existe sería conveniente generar.
c) Debe preprocesar los datos para asegurar que su modelo tenga la mejor entrada para el modelamiento y genere resultados confiables.
d) Debe utilizar únicamente los siguientes 3 algoritmos de ML: XGBoost, Light GBM y Redes Neuronales para realizar la función solicitada.
e) Debe hacer ingeniería y selección de características para el modelamiento, para lo cual debe justificar cada inclusión y eliminación de característica por medio de algún análisis estadístico, un gráfico o basada en lo que algún algoritmo indique que aporta o no para la función que usted está modelando.

f) Debe buscar alguna optimización de híper parámetros de cada algoritmo, para que su modelo alcance un mejor desempeño ya que su posición está en juego por estar en su periodo de prueba.
g) Una vez que haya encontrado su mejor modelo, deberá hacer la predicción para el dataset de evaluación que se ha entregado, el que contiene 7400 casos sin la etiqueta **net60mn**. Su predicción de tal etiqueta la debe entregar en un archivo .csv para obtener el puntaje de la competencia por los 14 puntos.

**F. Qué y cómo se evaluará este Proyecto** – Ponga mucha atención a esto!

a. Se evaluará la aplicación de metodología para el **EDA**.
b. Se evaluará la utilización de mejores prácticas para **pre-procesamiento**.
c. Se evaluará la **ingeniería de características y la justificación de la elección de variables** para el modelamiento. Si no hace selección de variables será penalizado con 7 puntos. (se restan).
d. Se evaluará la búsqueda de mejores parámetros para optimizar el desempeño del algoritmo.
e. También se evaluará la presentación del Jupyter Notebook (el .ipynb) en términos de comentarios y documentación de los pasos seguidos explicando el desarrollo. Este es uno de los entregables de este proyecto.
f. El otro entregable es el conjunto de archivos compuestos por un archivo .csv con la predicción realizada por el modelo que propondrá para la unidad de inversiones y por el modelo entrenado que propondrá al área de inversiones. (puntaje según lugar en desempeño del modelo: **1er lugar: 14 ptos, 2do: 12 ptos, 3ro: 10 ptos, 4to: 8 ptos y > 4to lugar: 4 ptos**).

**G. Entregables**

Para lo anterior se pide entrega, con la estructura del nombre indicado:

1. Archivo **HPC-GRUPO-nn.ipynb** con todo el código y los análisis. (donde nn=n° del grupo asignado), donde se deberá documentar lo que se hace con "markdown", con secciones claras de lo que se hace. Justificando inclusión de librería y documentando el código por bloques funcionales.

2. Documento en pdf (**HPC-GRUPO-nn.pdf)** con el siguiente contenido en secciones que el grupo de verá proponer:

   i) El análisis descriptivo de los datos, gráficos e **interpretación** de lo todo lo graficado y observado.
   ii) La **explicación de la estrategia** para abordar los aspectos técnicos del preprocesamiento.
   iii) **Justificación** de todas las **decisiones** de modelamiento.
   iv) La propuestas y justificaciones de grillas para la optimización de hiper-parámetros.
   v) La estrategia para responder al Banco Estelar.
   vi) Las recomendaciones para la selección de los clientes de alto perfil y potencial.
   vii) Los insights y conclusiones encontradas.

   Respecto del documento se evalua la formalidad presentada por ejemplo: tapa con grupo e integrantes, asignatura, tabla de contenidos, cabecera sobria y pie de página, ortografía, redacción, si está completo el contenido, explicación de los enfoques de desarrollo y estrategias propuestas y finalmente conclusiones.

3. **Modelo con mejor desempeño**: El mejor modelo de su trabajo se debe guardar en un archivo del tipo ".joblib" lo que se consigue con la librería joblib de Scikit-Learn.

4. **Archivo .csv con predicciones**: Cuando termine su modelo deberá usarlo para generar las predicciones del **dataset de evaluación**. Use df.to_csv, encoding='utf-8' y sin index df.to_csv(filaneme.csv, encoding='utf-8', index=False)

5. Buena documentación y presentación de su código en Jupiter (Python)

   i) Justificación del uso de librerías (la idea es que no haya en el código liberías que no se usan y que resulten de copias de código)
   ii) Nivel de documentación del código. (Orden y claridad del markdown usado para documentar el código)

# Tabla de Evaluación del Taller N°3

| Aspecto Evaluado | Puntaje Máximo |
|---|---|
| 1. EDA (profundidad, primeros insights, interpretaciones, uso del analizado, etc.) | 21 |
| 2. Pre-procesamiento (estrategias, uso de técnicas, como imputan, justificación de decisiones, etc.) | 21 |
| 3. Ingeniería y Selección de características | 28 |
| 4. Optimización hiperparámetros (criterios selección grilla, justificaciones, estrategia,etc) | 07 |
| 5. Modelamiento (uso de 3 algos ML y selección del mejor modelo, uso de métricas de desempeño, etc.) | 28 |
| 6. Ranking en predicción con dataset de evaluación (hpc_stellar_eval.csv) | 14 |
| 7. Código (documentado y librerías justificadas) | 07 |
| 8. Documento (completitud, ortografía, redacción, explicaciones y justificaciones de decisiones y estrategias + formalidad) | 14 |
| **Total Puntos** | 140 |

$$Nota\ Final\ Taller\ N°3 = \frac{1}{20}\sum puntaje\_aspecto_i$$

**Fecha de Entrega: Viernes 15-Diciembre-2023 18:00 hrs**

**BUEN DESARROLLO DE PROYECTO !!**

Información adicional:

- ✓ **Contexto del mercado chileno**: Chile ha tenido un crecimiento económico sostenido en la última década, con un PIB per cápita que supera los USD 25.000. Esto ha generado el surgimiento de una clase media-alta y alta con alto poder adquisitivo. Además, Chile tiene una relativa estabilidad política y regulatoria, lo que lo hace atractivo para la inversión extranjera. Por estos motivos, Chile se presenta como una oportunidad interesante para que un banco suizo de elite como Stellar ingrese ofreciendo productos financieros sofisticados a los segmentos de altos patrimonios.

- ✓ **Consejos sobre desafíos comunes:** Para el desbalance de clases, se puede usar oversampling o undersampling. Para valores faltantes, imputación o descarte según análisis de missingness. Usar técnicas como escalado para manejar outliers. Validación cruzada para evitar overfitting.

- ✓ **Guía sobre documentación en Jupyter Notebook**: Breve descripción de la función de cada bloque de código. Comentarios explicando decisiones críticas (elección de algoritmo, umbrales, etc). Interpretación de resultados y hallazgos en markdown. Referencia y justificación del uso de cada librería externa. Análisis del impacto de opciones evaluadas, pero no seleccionadas.

- ✓ **Posibles sesgos y mitigaciones**: Sesgos en características como género, edad o comuna. Evaluar y ajustar. Clases desbalanceadas pueden sesgar al modelo. Mitigar en preprocesamiento. Overfitting con pocos datos de entrenamiento. Usar regularización. Mala representatividad del conjunto de entrenamiento respecto a la población objetivo.

Aquí hay algunas sugerencias para la ingeniería de características que podrían ser útiles en el modelado predictivo:

- ✓ **Interacción Entre Edad y Nivel Educativo**: Crear una característica que combine la edad y el nivel educativo, como "edad_educacion", que podría ser el producto de la edad y los años de educación. Esto puede ayudar a identificar perfiles que tienen tanto experiencia como educación, lo cual puede ser un indicador de mayores ingresos.
- ✓ **Ponderación de la Ocupación**: Desarrollar una puntuación o ponderación para las ocupaciones basada en el nivel promedio de ingresos asociados con cada ocupación en el mercado laboral chileno. Esto podría ayudar a identificar ocupaciones que generalmente están asociadas con altos ingresos.
- ✓ **Ratio de Ganancia/Perdida de Capital**: Crear un ratio entre la ganancia y la pérdida de capital. Un ratio alto podría indicar una habilidad para invertir y gestionar el capital de manera efectiva, lo cual podría ser un indicativo de un cliente de alto perfil.
- ✓ **Indicador de Estabilidad Laboral**: Usar la combinación de edad, años en la ocupación actual y tipo de empresa para crear un indicador de estabilidad laboral. La estabilidad puede ser un buen indicador de ingresos constantes y, por ende, de un potencial cliente de alto perfil.
- ✓ **Índice de Inversión Potencial**: Basado en variables como la edad, el nivel educativo, la ocupación, y la ganancia/perdida de capital, crear un índice que estime la propensión de un individuo a invertir en productos financieros.
- ✓ **Agrupación Geográfica**: Utilizar la comuna de residencia para crear grupos geográficos basados en el nivel económico promedio o en la densidad de población de alto ingreso. Esto puede ayudar a identificar áreas con una mayor concentración de clientes potenciales de alto perfil.
- ✓ **Interacción entre Estado Civil y Relación con el Jefe de Hogar**: Combinar estas dos variables para crear una nueva característica que pueda indicar estabilidad familiar, lo cual puede estar correlacionado con la estabilidad financiera.