EAE3706 MACHINE LEARNING Y BIG DATA
1ER SEMESTRE 2022
INSTITUTO DE ECONOMÍA
PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

PROYECTO FINAL:
RESPONDER A UNA PREGUNTA DE INVESTIGACIÓN MEDIANTE LA APLICACIÓN DE MÉTODOS DE MACHINE LEARNING

- **Objetivo:** responder a una problemática real mediante la aplicación de modelos de Machine Learning sobre un set de datos.
- **Metodología:** aplicación del proceso de data science: formulación de la pregunta de investigación, aplicación de técnicas de manejo de bases de datos, entrenamiento de modelos predictores y comunicación de resultados.
- **Qué se evaluará:**
    - Describir la pregunta de investigación (ejemplo: ¿es posible mejorar las proyecciones de PIB utilizando modelos de machine learning?).
    - Entregar un contexto breve sobre el área de aplicación (ejemplo: describir concepto de crecimiento económico, importancia de predicciones en la formación de expectativas, etc.).
    - Describir el set de datos: origen, número de observaciones, columnas y sus tipos de datos, posibles sesgos asociados a la captura de los datos.
    - Realizar un análisis exploratorio de los datos: distribuciones de variables numéricas, probabilidad de ocurrencia de estados para variables categóricas, decisiones de imputación de missings y outliers, matriz de correlaciones entre variables numéricas.
    - Tomar decisiones de transformación de atributos, de ser necesario (normalización, reducción de dimensionalidad).
    - Entrenar al menos 2 modelos supervisados[1] de machine learning vistos en el curso, uno de cada uno de los siguientes grupos:
        - Grupo 1: Decision Trees, Random Forest, Naive Bayes
        - Grupo 2: Multilayer perceptron, Support Vector Machines
    - Comunicar los resultados de performance (accuracy, F1, ROC/AUC, tiempo de cómputo) de cada uno de modelos entrenados con distintas parametrizaciones, utilizando técnicas de visualización de datos.
    - Breve discusión sobre los resultados obtenidos y conclusiones: qué algoritmo lo hizo mejor y porqué, y su aporte a responder la pregunta de investigación.
- **Temas:** relacionados con Economía, Econometría, Finanzas, Políticas Públicas u otras áreas adyacentes.
- **Grupos de trabajo:** mínimo 2 y máximo 3 personas, sin asignación centralizada.
- **Fechas importantes:**
    - **15-Abril:** entrega de abstract del proyecto. Describir el problema a resolver y la fuente de datos a utilizar, entregando una captura de una muestra (10 observaciones) de la base de datos, donde se expliciten las columnas.
    - **27-Mayo:** entrega de avance (código en Jupyter notebook, extensión ".jpynb") con el trabajo de datos, descripción de variables y análisis exploratorio de los

[1] Eventualmente se puede aceptar un problema no supervisado, previa autorización. En esos casos se exigirá utilizar al menos un modelo determinístico (K-Means) y uno probabilístico (Gaussian Mixture Models).

datos. Se recomienda además incluir los modelos ya entrenados que se tengan hasta ese momento.

- **1-Julio:** entrega de código final ejecutado (Jupyter Notebook ".jpynb") y documento de máximo 6 páginas de contenido. Adicionalmente, entregar una muestra aleatoria de 1,000 observaciones de la base de datos utilizada con el fin de confirmar la ejecución del código en formato ".csv" (entregar dataset completo en caso de trabajar con menos de 1,000 observaciones).