Facultad de Ingenierías
Analitica de Datos 1
Ingeniería de Datos e IA
Prof. Cristian E Garcia.

Proyecto 1

Condiciones:

- *Subir el proyecto en formato* **pdf** *en la plataforma UAO-Virtual. Se debe entregar un informe y los archivos adicionales se deben eentregar como anexo. El informe no debe contener mas de 25 paginas incluida las imagenes y tablas.*
- *Es necesario incluir el código de R, Python o Julia. Mostrar los resultados a partir de tablas, gráficos o indicadores que les permita dar respuesta a los planteamientos.*
- *Deben interpretar los resultados obtenidos en cada situación de acuerdo al contexto.*
- *Realizar la actividad en los grupos Definidos anteriormente.*
- *Subir la tarea en un archivo* **.ZIP,** *donde contenga el informe en pdf y los ejecutables de códigos que haya utilizado para el desarrollo del Proyecto.*

Situación 1 (0.75 Pts)

Se dispone de un conjunto de **datos (link)** multivariados ($n = 200, p = 3000$) proveniente de la evaluación de un Modelo de Lenguaje de Gran Escala (LLM), NLA-7B. Las n observaciones corresponden a *prompts* (consultas) únicos, mientras que las p variables representan un conjunto heterogéneo de métricas (activaciones internas, evaluaciones humanas, y medidas de rendimiento derivadas). El dataset exhibe un problema de alta dimensionalidad ($p \gg n.$), lo cual invalida la aplicación directa de métodos multivariados clásicos.

El estudio persigue un doble objetivo:

1. Validar psicométricamente un sub-instrumento de 20 ítems de evaluación humana, cuya estructura teórica postula cuatro constructos latentes (Calidad, Seguridad, Creatividad, Sesgo).
2. Implementar una estrategia de **Análisis Factorial-Cluster (Factor-Cluster Analysis)** para identificar perfiles tipológicos (clústeres) de los *prompts* basados en la totalidad de las $p = 3000$ variables, superando el desafío $p \gg n$.

Validación de la Estructura del Instrumento:

- Aísle el subconjunto de $p = 20$ variables correspondientes a la evaluación humana.
- Especifique el modelo de medición de cuatro factores (Calidad, Seguridad, Creatividad, Sesgo) basado en la hipótesis teórica.

- Estime el modelo y evalúe la bondad de ajuste (e.g., RMSEA, CFI, SRMR, χ2). Concluya sobre la validez de constructo del instrumento.

Reducción de Dimensionalidad y Extracción de Características (EFA/PC):

- Considerando la matriz de datos completa (200×3000), discuta formalmente las implicaciones de la condición $p \gg n$ sobre la estimación de la matriz de covarianza (singularidad) y la inestabilidad de métodos como Máxima Verosimilitud.
- Implemente un Análisis Factorial Exploratorio utilizando el método de **Componentes Principales (PC)** para la extracción factorial, justificando su idoneidad y estabilidad computacional en este contexto.
- Determine el número de factores/componentes (m) a retener, utilizando el **Gráfico de Sedimentación (Scree Plot)** y el criterio de varianza explicada acumulada.
- Rote la solución factorial (e.g., Varimax) y genere la matriz de **puntuaciones factoriales** $(n \times m)$, la cual servirá como el nuevo espacio de características de baja dimensión.

Identificación de Tipologías:

- Utilizando la matriz de puntuaciones factoriales $(200 \times m)$ como entrada, justifique la superioridad de este enfoque (Análisis Factorial-Cluster) frente al clustering sobre el espacio original de $p = 3000$.
- Aplique un algoritmo de particionamiento (K-Means) y un método jerárquico aglomerativo (e.g., Método de Ward) sobre el espacio m-dimensional.
- Determine el número óptimo de clústeres (k) (vía Coeficiente de Silueta y análisis del dendrograma) y seleccione la solución de particionamiento más robusta.

Caracterización e Interpretación de Perfiles:

- Para la solución de k clústeres seleccionada, realice el perfilado de cada conglomerado. Calcule los centroides (medias) de las puntuaciones factoriales (m dimensiones) para cada clúster.
- Interprete el significado conceptual de cada clúster en el contexto del problema (e.g., "Perfil 1: Prompts de alta creatividad y alto riesgo de sesgo", "Perfil 2: Prompts de alta calidad factual y baja complejidad interna").
- Sintetice los hallazgos en un informe, discutiendo las implicaciones de estas tipologías para el *fine-tuning* y la evaluación del modelo NLA-7B.

Situación 2 (0.75 Pts)

Una agencia de gestión agrícola necesita realizar una evaluación rápida del estado de la cobertura terrestre en una región vulnerable después de un evento climático extremo (e.g., sequía o inundación). No poseen etiquetas actualizadas (ground truth) de la zona, pero tienen acceso a miles de imágenes satelitales de alta resolución de la región.

El objetivo es utilizar **clustering** para agrupar estas imágenes en categorías (clústeres) que representen las diferentes *tipologías de cobertura terrestre* (e.g., cultivo saludable, bosque denso, suelo desnudo, zona inundada). Se hipotetiza que los clústeres que representen "suelo

desnudo", "zona industrial" o "agua estancada" son indicadores de *zonas agroecológicas dañadas* o en estrés.

Fuente de Datos

- **Dataset: EuroSAT**: Land Use and Land Cover Classification with Sentinel-2
- **Descripción:** Este conjunto de datos contiene 27,000 imágenes satelitales (64x64 píxeles) tomadas por el satélite Sentinel-2.
- **Clases (Etiquetas):** Las imágenes están etiquetadas con 10 clases que usaremos *solo para la validación final*. Las clases incluyen:
  - `AnnualCrop` (Cultivo Anual)
  - `PermanentCrop` (Cultivo Permanente)
  - Pasture (Pasto)
  - `Forest` (Bosque)
  - HerbaceousVegetation (Vegetación Herbácea)
  - `Industrial` (Zona Industrial)
  - `Residential` (Zona Residencial)
  - `SeaLake` (Mar/Lago)
  - River (Río)

Modelado de Clustering

Aplique algoritmos de clustering sobre los datos reducidos (n×m).

1. K-Means:
   - Determine el número óptimo de clústeres (k) utilizando el **método del codo (SSE)** y, más importante, el **Coeficiente de Silueta**.
   - Ajuste el modelo K-Means con la k óptima (e.g., k=8 a 12).
2. DBSCAN:
   - Aplique DBSCAN. Este método es robusto a los *outliers* y no requiere especificar k.
   - Justifique la elección de sus hiperparámetros (`epsilon` y `min_samples`).
   - ¿Cuántos clústeres encuentra DBSCAN? ¿Identifica muchas imágenes como "ruido" (anomalías)?

Evaluación y Perfilado (La Conexión Agroecológica)

1. **Validación Interna:** Reporte el **Coeficiente de Silueta** de su mejor modelo de clustering.

2. Validación Externa (Usando las Etiquetas Ocultas):
   - Ahora, y solo para validar, utilice las etiquetas reales del dataset EuroSAT (e.g., `AnnualCrop`, `Industrial`, etc.).
   - Genere una **matriz de confusión** cruzando sus k clústeres (Cluster 1, Cluster 2...) con las 10 clases reales.
   - Calcule métricas de validación externa como el **Adjusted Rand Index (ARI)** o **Normalized Mutual Information (NMI)** para medir qué tan bien su segmentación no supervisada coincide con la realidad.
3. Perfilado y Conclusión:
   - Inspeccione la matriz de confusión. ¿Qué clases componen cada clúster?

- **Conclusión:** Identifique los clústeres que representan "zonas agroecológicas dañadas" o en estrés. ¿Son estos los clústeres que agruparon las imágenes de `Industrial`, `SeaLake` o `Residential`? ¿Creó el algoritmo un clúster separado para los `Cultivos` vs. `Bosques`?

Situación 3 (3.5 Pts)

El conjunto de datos *Single-Family Fixed-Rate Loan Performance Data* ofrece una visión longitudinal y granular del desempeño de préstamos hipotecarios en Estados Unidos. La alta dimensionalidad y la interdependencia temporal de estas variables (originación, desempeño mensual, características del prestatario) presentan un desafío analítico. La identificación de patrones subyacentes, más allá de los indicadores de riesgo tradicionales, es fundamental para una gestión de cartera proactiva, una tarificación (pricing) precisa y el diseño de políticas de originación robustas.

A pesar de la disponibilidad de indicadores de riesgo convencionales (como el puntaje FICO o la relación Préstamo-Valor), estos suelen ser estáticos y fallan al capturar la dinámica evolutiva del comportamiento del prestatario y las condiciones macroeconómicas. Existe la necesidad de desarrollar un marco analítico capaz de:

1. Procesar y armonizar datos masivos distribuidos en múltiples periodos anuales.
2. Identificar variables latentes (constructos no observables directamente) que expliquen la variabilidad del desempeño crediticio.
3. Implementar modelos de segmentación que integren tanto el rigor estadístico del Análisis Factorial como la capacidad de representación no lineal del Deep Learning.

Objetivo General

Diseñar e implementar una solución analítica de alto rendimiento, basada en computación en la nube y técnicas de aprendizaje profundo, para identificar, modelar y caracterizar estructuras latentes que permitan una segmentación avanzada del riesgo en un portafolio hipotecario a gran escala.

Objetivos Específicos

- Construir un panel analítico longitudinal utilizando infraestructuras de *Cloud Computing* y procesamiento distribuido la para la información.
- Ejecutar un Análisis Factorial Exploratorio (AFE) y Confirmatorio (AFC) integrando técnicas de reducción de dimensionalidad para datos mixtos y autoencoders.
- Desplegar arquitecturas de Deep Learning para capturar la interdependencia temporal de los flujos de pago.
- Desarrollar un modelo de agrupamiento (*clustering*) basado en los espacios latentes obtenidos para definir perfiles de riesgo interpretables.

Fases del Estudio (Metodología)

Ejecutar secuencialmente las siguientes fases analíticas:

1. Construcción del Panel Analítico: Integrar y preparar las diversas tablas del conjunto de datos (originación, desempeño mensual, atributos del prestatario e inmueble) para construir un panel analítico longitudinal y coherente. Este panel debe ser apto para el estudio conjunto de variables de riesgo, capacidad de pago y desempeño histórico.

2. Extracción de Componentes Latentes: Aplicar un método de **Análisis Latentes** adecuado para los datos. El objetivo es extraer componentes *multivista* que capturen la información compartida entre distintos grupos de variables (p. ej., dominios del préstamo, del prestatario y del comportamiento mensual). Se busca que estas fuentes latentes expliquen la heterogeneidad multivariada del portafolio de manera compacta.

3. Evaluación e Interpretación de Componentes: Evaluar y contrastar los componentes extraídos con los indicadores tradicionales de riesgo crediticio. El análisis debe centrarse en la interpretabilidad, estabilidad y relevancia de los componentes para caracterizar patrones de morosidad, prepago, incumplimiento y amortización.

4. Segmentación Basada en Características Latentes: Utilizar las proyecciones (scores) de los componentes independientes obtenidos como base para un proceso de **Análisis de Conglomerados**. Se deberán explorar técnicas como k-means, Gaussian Mixture Models (GMM) o clustering jerárquico para segmentar los préstamos en grupos homogéneos.

5. Caracterización de Perfiles de Riesgo: Interpretar y validar los conglomerados resultantes. La caracterización debe realizarse en términos de perfiles de riesgo, rasgos financieros del prestatario y propiedades estructurales del préstamo.

Requerimientos Técnicos

Fase I: Arquitectura y Cloud Computing

Proponer una arquitectura de datos que garantice la escalabilidad. Se requiere el uso de:

- **Almacenamiento:** Formatos de archivo optimizados (Parquet/Avro) en servicios de almacenamiento de objetos.
- **Computación:** Despliegue de clústeres para procesamiento en memoria (ej. Spark, Dask).

Fase II: Análisis Multivariado y Latente

Aplicar:

1. **Análisis de Componentes Latentes:** Identificación de dimensiones no observadas que correlacionen atributos del prestatario, el inmueble y el mercado.
2. **Validación de Estructura:** Aplicar AFC para contrastar las hipótesis de las dimensiones extraídas con la teoría financiera del riesgo.

Fase III: Componente de Deep Learning e IA

Integración de una solución de Deep Learning para la generación de *embeddings* de riesgo:

- **Autoencoders (VAE):** Para la representación compacta y no lineal de la alta dimensionalidad.
- **Frameworks:** Uso de librerías avanzadas para el entrenamiento distribuido.

Resultados Esperados y Evaluación

El proyecto culminará con un informe que debe incluir:

- **Arquitectura de Solución:** Diagrama detallado del flujo de datos en la nube y los modelos de IA empleados.
- **Análisis:** Discusión sobre la estabilidad de los componentes latentes frente a choques económicos.
- **Impacto de Negocio:** Discusión dobre cómo la segmentación obtenida optimiza la tarificación (*pricing*) y el monitoreo de cartera proactivo.

Criterios de Evaluación: Integración tecnológica, validación estadística de los modelos latentes, eficiencia computacional de la solución y profundidad en la interpretación de los perfiles de riesgo identificados.

Reubrica

La rubrica del proyrecto se encuetra adjunta en el siguiente **link**