Taller 1: Preprocesamiento y Aprendizaje No Supervisado
Ciencia de Datos
Profesor: Gabriel Jara
Primer Semestre 2025

El objetivo del Taller 1 es desarrollar actividades de preprocesamiento y exploración de datos, aplicando técnicas de limpieza y preparación de datos, así como también algoritmos de aprendizaje no supervisado.

Cada año las universidades de Chile reciben cientos de miles de postulaciones, a través del sistema centralizado de admisión, que gestiona DEMRE. Se le provee de tres archivos separados por coma con datos provenientes del proceso 2025. El archivo B corresponde a todas las inscripciones para rendir las pruebas de selección; el archivo C corresponde a los resultados de los postulantes en las diversas pruebas; el archivo D corresponde a las postulaciones a carreras de las distintas universidades. Cada uno de estos archivos SCV es acompañado por una planilla Excel con los códigos y datos complementarios, lo cual permite, entre otras cosas, identificar con precisión cada universidad y carrera en el set de datos.

Cabe señalar que los postulantes están identificados a través de un ID único, lo que permite consolidar la información de un postulante a través de los tres archivos de datos. Dicho ID ha sido enmascarado por DEMRE, de modo que no viola la privacidad de las personas, a la vez que permite hacer análisis sobre los tres archivos de manera consistente.

Los estudiantes de la asignatura Ciencia de Datos pueden realizar el taller en parejas o individualmente. Se presenta a continuación una serie de actividades que deberán realizar, junto al puntaje asignado. Para cada actividad, se solicita que ingrese una descripción de lo realizado junto a evidencia que puede incluir gráficos y salidas de código. El informe debe ser entregado en formato **Jupyter Notebook** (ipynb). Se requiere que los notebooks vengan **ya ejecutados** (aunque es probable que se vuelvan a ejecutar durante la evaluación).

Las actividades suman 120 puntos, pero la nota máxima será 100.

Atención: Instrucciones importantes

- El taller se debe desarrollar en Jupyter Notebook, debe entregar el archivo ***ipynb***. Su notebook debe poder cargar los archivos csv con dados sin intervención adicional (más allá de asegurarse que esté en la misma carpeta del notebook). Es decir, toda edición que se le pide hacer **debe hacerla usando Python desde el notebook.**
- Ponga nombre(s) de los autores en el encabezado del Jupyter Notebook y también en el nombre del archivo (ej: “taller_1_Gabriel_Jara_Perico_delosPalotes.ipynb”). Haga esto desde un principio, que no se olvide.
- Entregue su Jupyter Notebook **ya ejecutado**. Esto quiere decir que se espera que los resultados que genera su código ya vengan a la vista en la versión que entregue. Sin desmedro de lo anterior, considere que su Jupyter podría ser ejecutado durante el proceso de evaluación, y se espera que su código corra sin errores.

- Suba al portal de entrega en aula su Jupyter Notebook (archivo ipynb) a más tardar el **domingo 04 de mayo a las 23:59**.
- Si hizo el trabajo en pareja, sólo uno de los integrantes debe subir el archivo (que contiene el nombre de ambos).

Actividad 1: Descripción del Set de Datos de Postulantes (10 puntos)

Debe usar los tres archivos provistos para construir un set de datos que consolide toda la información relevante para analizar a los Postulantes. Deberá tomar algunas decisiones, por ejemplo, respecto a cómo consolidar la información de postulaciones dado que cada postulante tiene una a muchas postulaciones. Puede hacerlo de distintas formas, pero sea claro en explicar cómo construye su set de datos.

Explorando los datos notará que la muchos de ellos están expresados numéricamente, pero en realidad corresponden a una codificación, por ejemplo: número de comuna. Otros, en cambio, sí son atributos numéricos propiamente tal. De estos últimos seleccione alguno para analizar, identifique rango, tendencia central, dispersión, presente alguna gráfica que ilustre el comportamiento de dicha variable.

Seleccione también dos atributos categóricos, uno nominal y otro ordinal. Describa cada uno de estos dos atributos presentando gráfica que ilustre el comportamiento de dicha variable.

Actividad 2: Calidad y Limpieza de Datos (30 puntos)

Realice al menos **cuatro tareas de evaluación de calidad y/o limpieza de datos**. Procure que sean actividades diferentes unas de otras, no la misma tarea aplicada a distintos atributos u observaciones. Identifique problemas que puedan afectar el análisis o el correcto funcionamiento de los algoritmos, explique cómo los soluciona y qué impacto podrían tener en los resultados.

Cada tarea de limpieza realizada debe tener su celda con código y explicación correspondiente.

Si decide imputar valores faltantes, por favor considere para cada variable qué significa un valor vacío. En algunos casos la ausencia de valor representa que el postulante no rindió una prueba (por ejemplo), y en ese caso usar la media como valor de imputación distorsionaría la información presente en los datos.

Actividad 3: Clustering (40 puntos)

Aplique al menos **dos algoritmos de Clustering distintos** para agrupar a los postulantes. Utilice técnica de reducción de dimensionalidad en al menos uno de sus experimentos. Para cada algoritmo, pruebe al menos **dos configuraciones de parámetros** o modificaciones al set de datos, para un total de al menos **cuatro agrupamientos distintos**.

- Comente los resultados de cada agrupamiento.
- Al utilizar reducción de dimensionalidad, explique qué diferencias se observan antes y después de aplicarla.
- Compare los resultados obtenidos al cambiar los parámetros o al modificar los datos.

Nota: Puede reducir el set de datos a un subconjunto de atributos o a una submuestra. Considere que algunos algoritmos pueden tardar un tiempo considerable si se usan todos los datos. Con K-Means no debería tener problema en usar todos los atributos, pero con otros algoritmos el tiempo de espera se podría volver considerable.

Actividad 4: Reglas de Asociación (40 puntos)

Los datos incluyen identificadores de postulantes, carreras y Universidades, lo que permite identificar postulaciones a distintas carreras (y universidad) de cada postulante. Construya un set de datos adecuado para extraer reglas de asociación sobre la postulación a las carreras (y universidad).

Las reglas de asociación que se requiere obtener deben relacionar: una carrera (universidad) en el antecedente => una o más carreras (universidades) en el consecuente.

Aplique un algoritmo de reglas de asociación (por ejemplo, **Apriori**) y genere al menos **cinco reglas** de la forma solicitada. Comente los resultados y especule sobre posibles explicaciones para las asociaciones encontradas, utilizando al menos **dos métricas** para evaluar la relevancia de las reglas.

Un ejemplo del tipo de regla que sería válida, que vincula postulación a carreras:

- Ingeniería en Informática (UTFSM) **=>** Ingeniería Electrónica (UTFSM), Programación (UCV)

Ejemplos de reglas NO VÁLIDA, que resultarían si no se ha preprocesado adecuadamente el set de datos:

- Provincia=Valparaíso **=>** Comuna=Valparaíso
- Régimen=Femenino **=>** Género=Femenino

Nota: En este taller se espera que se especule un poco en la interpretación de sus resultados. Comente sus teorías respecto a qué al porqué de los agrupamientos y reglas que obtenga. Procure dejar claro, eso sí, lo que es observación objetiva y lo que es interpretación de su parte.

Rúbrica

Criterio	0%	25%	50%	75%	100%
Descripción de Datos	No entrega descripción o es incorrecta	Entrega parcialmente la descripción	Describe mayoritariamente bien, pero con omisiones	Buena descripción, sin errores importantes	Completa y detallada descripción con análisis
Calidad y Limpieza	No realiza evaluación o está incorrecta	Realiza sólo una tarea o está incompleta	Dos tareas completas, pero con omisiones o errores	Dos tareas correctas, sin errores relevantes	Evaluación completa con análisis adicional
Clustering y Reducción	No aplica clustering o es incorrecto	Realiza parcialmente el clustering	Clustering correcto, pero sin usar reducción o análisis incompleto	Realiza clustering y reducción, pero sin análisis detallado	Clustering con reducción y análisis completo
Reglas de Asociación	No genera reglas.	Genera reglas incorrectas (no válidas).	Genera reglas válidas, sin análisis sustentado en métricas.	Genera reglas válidas, con análisis parcial.	Reglas bien analizadas con interpretación clara basada en métricas.

La nota máxima será **100**, aunque las actividades sumen más puntos. Si se sacó un 120, felicidades tiene un 100.