Laboratorio 1 - Análisis Exploratorio de Datos
Estadística Computacional

Universidad Técnica Federico Santa María - Campus Casa Central
Departamento de Informática

18 de agosto del 2025

Dra. Carolina Saavedra
<carolina.saavedra@usm.cl>

Nicolás Armijo
<nicolas.armijoc@usm.cl>

Diego Bahamondes
<diego.bahamondes@usm.cl>

Diego Concha
<diego.conchab@usm.cl>

1. Introducción

El laboratorio de estadística computacional busca reforzar los conocimientos adquiridos en el ramo de estadística mediante un proyecto. El objetivo es ofrecer una experiencia valiosa para el estudiante, permitiéndole desarrollar habilidades prácticas que son muy valoradas en el mercado laboral, además de ser la primera puerta para el mundo del Data Science. La capacidad para analizar datos, tomar decisiones informadas, y comunicar resultados de manera clara y efectiva son habilidades esenciales para una multitud de carreras.

Imaginen que junto a su grupo se encuentran trabajando como científicos y científicas de datos en una organización reconocida a nivel mundial. Por el prestigio de tal organización,

se les asigna la libertad de investigar y explorar cualquier fenómeno del mundo real: como educación, economía, marketing, el crimen, salud, entre otros.

Esta investigación es la que se llevará a cabo durante el resto del semestre. Para esta primera instancia, su misión es **formular una pregunta o hipótesis**, la que buscarán responder durante las siguientes experiencias.

2. Desarrollo

2.1. Conjunto de datos: Elige tu terreno

El primer paso es seleccionar un conjunto de datos que despierte la curiosidad del grupo. El conjunto de datos seleccionado debe ser aprobado por su tutor, donde debe cumplir los requisitos:

- Debe contar con un **mínimo** de 200 observaciones.
- Debe contener por lo menos **2 variable de naturaleza categórica**, y **2 variable de naturaleza continua**. (Investigar sobre los distintos tipos de datos)
- Especificar claramente la motivación para seleccionar el dataset.

Algunas páginas útiles para encontrar datasets son:

- Kaggle datasets: kaggle.com/datasets
- Datasets del gobierno de Chile: gob.cl/datasets
- Datasets subreddit: r/Datasets
- UCI Irvine Machine Learning Repository: uci.edu/datasets

2.2. Análisis Exploratorio de Datos

El análisis exploratorio de datos es el paso inical de cualquier proyecto de Data Science. Por lo general, se realizan una serie de pasos que son aplicables para la gran mayoría de datasets. Se les pide investigar en qué consisten las siguientes tareas para luego aplicarlas al conjunto de datos escogido previamente:

1. Limpieza de datos: Tratar valores faltantes, codificaciones extrañas y posibles errores.
2. Descripción de las variables y datos: Identificar los tipos, calcular estadísticos para saber cómo se distribuyen, rangos, e identificar valores atípicos o también conocidos como outliers.
3. Visualizaciones: Realizar gráficos que ayuden a comprender tendencias, relaciones, diferencias, agrupamientos y otros comportamientos presentes en los datos.

4. Observar con lupa: Realizarse las siguientes preguntas ¿qué patrones saltan a la vista? ¿qué cosas no esperabas? ¿hay alguna relación interesante? *La profundidad del análisis es muy importante para ser capaces de plantear una buena pregunta o hipótesis de investigación.*

Además, deberán escoger 4 variables, 2 principales y 2 secundarias, con el fin de reducir el área que abarcarán en futuras entregas. Se recomienda que intenten alinear esta selección a su pregunta o hipótesis, donde deberán justificar sus decisiones basándose en resultados encontrados o sospechas que el grupo considere pertinentes.

2.3. Pregunta de Investigación o Hipótesis

Una vez realizada la exploración del conjunto de datos seleccionado, se deberá definir una pregunta de investigación o hipótesis estadística inicial. Una buena pregunta no solo refleja curiosidad, sino que debe ser medible y respondible con el dataset que se posee.

En futuras entregas se realizará inferencia y pruebas de hipótesis, por lo que la pregunta debe permitir un análisis estadístico concreto.

Las características de una buena pregunta pueden resumirse de la siguiente forma:

→ Es **clara**: No deja espacio a interpretaciones ambiguas.

→ Es **específica**: No es demasiado amplia ni difusa.

→ Es **medible**: Las variables están disponibles y se pueden cuantificar.

→ Es **abordable**: No requiere información extra.

→ Es **relevante**: Tiene sentido para el dataset.

→ Es **verificable**: Puede ser confirmada o refutada con un método estadístico.

Una buen camino es detectar algo interesante en en análisis exploratorio y transformarlo en una pregunta sin asumir la respuesta. Esta pregunta debe necesitar un test estadístico para poder responderse.

3. Sobre el desarrollo

El archivo de datos debe ser analizado con el lenguaje Python, junto a alguna librería de análisis de datos, tal como puede ser Pandas. El uso de esta librería no es estrictamente obligatorio, pero si altamente recomendado, debido a que dispone de funciones para muchas de las tareas que se pedirán realizar. Además, será necesario usar alguna librería de visualización de datos, por ejemplo, Matplotlib o Seaborn.

Se deberá realizar el análisis en un Jupyter Notebook, que permite mezclar código, visualización de datos y anotaciones en un mismo archivo.

Una opción muy interesante es desarrollar el análisis usando Google Colab (`https://colab.research.google.com`), el cual permite programar en el navegador usando Python en formato de Jupyter Notebook, ofreciendo todas las librerías ya instaladas, y permitiéndoles compartir el código fácilmente. Además, el sitio permite descargar el archivo en formato (.ipynb).

Finalmente, se deberán presentar el análisis y resultados encontrados en horarios a coordinar con los tutores.

4. Condiciones de entrega

La entrega deberá realizarse en AULA, en la sección "LEC -Proyecto/Entrega 1", con plazo máximo el **Domingo 31 de agosto a las 23:55 hrs**. Solo el encargado del grupo deberá realizar la entrega, la cual constará de un archivo comprimido .rar, .zip o .tar.gz, que deberá ser llamado .^entrega1-grupoX"(reemplazando el grupo según corresponda) donde será necesario que contenga:

1. El material que será utilizado para la presentación.
2. El código desarrollo en un archivo .ipynb.
3. El archivo del conjunto de datos seleccionado.

Se evaluará tanto la presentación y el desarrollo del Jupyter Notebook tal como se especifica en las reglas del curso y bajo las siguientes rúbricas.

Rúbricas

Rúbrica Jupyter

Dimensión evaluada	**Descripción**	**Puntaje**
Elección del dataset y contextualización	Se seleccionó un dataset relevante y bien justificado. Se explicó su contexto, origen y variables.	15 pts
Calidad del análisis exploratorio (EDA)	Se aplicaron técnicas apropiadas de limpieza, resumen estadístico y visualización de datos.	25 pts
Interpretación y comunicación de hallazgos	Se comentan y explican adecuadamente los resultados del EDA. Se identifican patrones y relaciones.	20 pts
Formulación de la pregunta o hipótesis inicial	Se propone una pregunta clara, cuantificable y coherente con el análisis exploratorio realizado.	15 pts
Presentación del informe (estructura y claridad)	El informe está bien organizado, con narrativa fluida, lenguaje técnico adecuado y visualmente claro.	15 pts
Código y reproducibilidad	El código está completo, comentado y permite reproducir los análisis realizados.	10 pts
Total		**100 pts**

Rúbrica Presentaciones

Criterio	Deficiente 0-29	Insuficiente 30-45	Suficiente 46-69	Bueno 70-89	Excelente 90-100
1.- Expresión oral, no verbal y claridad	La comunicación es confusa, desorganizada y difícil de entender. El grupo interrumpe constantemente, habla en voz baja y muestra falta de respeto y formalidad (p. ej., no respetan turnos o permanecen sentados mientras otro presenta).	La comunicación es poco clara y carece de coherencia. La actitud es irregular, con distracciones frecuentes, falta de respeto ocasional y comportamientos informales (p. ej., uso de celulares o poca participación).	La comunicación es generalmente clara pero con vacilaciones o imprecisiones. La actitud es mayormente respetuosa, aunque con inconsistencias y falta de cohesión en algunos momentos (p. ej., falta de atención o apoyo entre integrantes).	La comunicación es clara, organizada y coherente. La actitud es formal y respetuosa en la mayoría del tiempo, mostrando buena interacción y apoyo, aunque puede presentar pequeños descuidos (p. ej., algunos integrantes no mantienen contacto visual o formalidad constante).	La comunicación es clara, fluida y bien estructurada. La actitud es profesional, respetuosa y formal durante toda la presentación, con excelente trabajo en equipo, apoyo mutuo y cumplimiento estricto de las normas (p. ej., todos participan activamente, respetan turnos y mantienen una postura profesional).
2.- Orden, coherencia y estructura de la presentación	Presentación confusa y mal estructurada. Los temas se presentan de forma desordenada, sin una secuencia lógica. Las ideas no están conectadas entre sí, dificultando significativamente la comprensión del contenido. No se percibe planificación ni coordinación entre los expositores.	Presentación con una estructura débil y poco clara. Existen cambios de tema abruptos y falta de cohesión entre las ideas. Hay poca fluidez en la exposición, por lo que se debe hacer un esfuerzo constante para seguir el hilo conductor. La conexión entre secciones es deficiente o ausente.	Presentación con estructura reconocible, aunque con desorganización parcial. Las ideas están relacionadas pero las transiciones son poco claras, y en algunos momentos se pierde el foco del contenido. La coordinación es aceptable.	Presentación bien estructurada y coherente. Las ideas se presentan en orden lógico, con transiciones comprensibles. La exposición fluye en general, aunque pueden existir detalles menores por mejorar en ritmo o coordinación grupal.	Presentación muy bien organizada, clara y coherente. Las ideas fluyen naturalmente con transiciones fluidas y bien integradas. Existe una narrativa efectiva desde el inicio hasta el cierre, demostrando preparación y coordinación destacadas.
3.- Congruencia con Jupyter Notebook	La presentación no sigue el contenido del notebook, mostrando resultados que no están u omitiendo partes importantes. Hay contradicciones relevantes entre lo expuesto y el trabajo realizado, lo que indica poca revisión o comprensión del material entregado.	Existen múltiples inconsistencias o desajustes con los resultados del notebook. Se omiten partes clave del análisis. La presentación no refleja fielmente lo obtenido en el trabajo, lo que afecta negativamente el contenido expuesto.	La presentación refleja el contenido general del notebook, pero con algunas omisiones o imprecisiones. Puede haber errores menores en la forma de comunicar los resultados o falta de claridad en cómo se conectan con lo mostrado. La coherencia con el análisis entregado es aceptable, aunque podría mejorar.	La presentación es mayormente coherente con los resultados obtenidos en el Jupyter Notebook. Se muestran y explican correctamente los análisis relevantes, aunque pueden existir discrepancias puntuales o detalles omitidos. La interpretación se alinea con lo trabajado, pero podría ser más precisa en algunos aspectos.	La presentación es completamente coherente con el contenido y los resultados del Jupyter Notebook. Se muestran todos los análisis y conclusiones de manera fiel, respetando su contexto y alcance. El equipo demuestra un manejo sólido de lo realizado, asegurando consistencia total entre el trabajo técnico y la exposición oral.
4.- Correctitud Técnica y manejo de los conceptos	El grupo no demuestra comprensión de los conceptos clave del trabajo. Confunden términos, aplican mal los contenidos o no pueden explicarlos adecuadamente (p. ej., definen conceptos incorrectamente o los usan fuera de contexto).	El grupo tiene una comprensión muy limitada de los conceptos. Cometen errores importantes o dan explicaciones poco claras y superficiales (p. ej., mezclan ideas sin fundamento o aplican definiciones de forma imprecisa).	El grupo maneja los conceptos generales del tema, pero con vacilaciones, omisiones o explicaciones incompletas. Demuestran comprensión parcial (p. ej., reconocen conceptos pero no los relacionan correctamente con el análisis).	El grupo demuestra un buen manejo de los conceptos y sabe aplicarlos correctamente. Aunque hay pequeñas imprecisiones o falta de profundidad en algunos puntos, la comprensión general es clara y adecuada.	El grupo demuestra dominio completo de los conceptos, los explica con claridad, los relaciona con el trabajo realizado y los aplica correctamente en el análisis (p. ej., justifican decisiones usando teoría y términos con propiedad).
5.- Conclusiones y análisis	El grupo no presenta conclusiones o estas no tienen relación con el trabajo realizado. No hay análisis o se limita a repetir información sin reflexión ni sentido crítico.	El grupo presenta conclusiones muy generales, poco fundamentadas o con errores importantes. El análisis es superficial y no aporta valor al trabajo	El grupo entrega conclusiones acordes al trabajo, pero con poca profundidad o sin vincular completamente con el análisis realizado. El cierre es funcional, pero puede mejorar en coherencia y desarrollo crítico.	El grupo formula conclusiones claras y coherentes con el análisis realizado. El análisis muestra interpretación adecuada, aunque podría ser más profundo o propositivo	El grupo presenta conclusiones bien elaboradas, claramente vinculadas al trabajo y basadas en evidencia. El análisis es profundo, reflexivo y crítico, demostrando una comprensión global del tema y capacidad de síntesis
6.- Respuestas a preguntas del equipo de ayudantes	No responde a las preguntas planteadas por el equipo de ayudantes, mostrando falta de preparación o comprensión del tema.	Responde solo parcialmente o de forma confusa. Algunas respuestas son incorrectas o incompletas, evidenciando comprensión limitada y poca seguridad.	Responde correctamente a la mayoría de las preguntas, aunque con explicaciones poco claras o superficiales. Muestra comprensión general, pero con falta de profundidad o precisión.	Responde correctamente a las preguntas, demostrando buen dominio del tema, aunque puede omitir detalles o no profundizar lo suficiente en algunas respuestas.	Responde de manera completa, precisa y clara a todas las preguntas, incluyendo detalles relevantes, explicaciones bien fundamentadas y demostrando un conocimiento profundo y seguro del tema.