Programa de Estudios

MINERÍA DE DATOS

Código	: DOA_0006_CA01
Periodo Académico	: Segundo semestre 2025
Créditos	: 6
Requisito	: Ninguno
Horario	: Jueves de 16.00 a 18.15 hrs. (GMT -04.00, Santiago, Chile)
Profesor(es)	: Bastián González-Bustamante (bastian.gonzalez.b@mail.udp.cl)
Ayudante(s)	:

Última actualización: 7 de agosto de 2025

I. DESCRIPCIÓN

En este curso de minería de datos los estudiantes aprenderán a procesar grandes cantidades de datos para tomar decisiones informadas en diferentes tipos de organizaciones y llevar a cabo investigaciones en ciencias sociales. La minería de datos es un proceso fundamental para obtener conocimientos valiosos a partir de los datos que nos rodean. Al analizar y modelar estos datos, podemos identificar patrones y tendencias que nos permiten tomar decisiones más efectivas y mejorar procesos en áreas como el marketing, la gestión de recursos humanos y la investigación social en general.

Durante este curso, los estudiantes aprenderán conceptos básicos sobre minería de datos relacionados con la limpieza y el procesamiento de información, técnicas de imputación, detección de *outliers*, control de versiones con Git y análisis exploratorio y de clústeres. Además, se profundizará en el modelamiento predictivo utilizando algoritmos y árboles de decisión, técnicas de ensamblado de modelos y una breve aproximación al procesamiento de lenguaje natural (*Natural Language Processing*, NLP) y grandes modelos de lenguaje (*Large Language Models*, LLMs) para la clasificación y análisis de datos.

Se empleará la metodología de Aprendizaje Basado en Investigación (ABR), donde los estudiantes, a través de investigaciones individuales, abordarán un problema central. Al finalizar este curso, los estudiantes habrán desarrollado las habilidades necesarias para procesar y analizar grandes volúmenes de información y estarán capacitados para aplicar estos conocimientos en distintos contextos.

II. OBJETIVOS

Al final del curso se espera que los y las estudiantes sean capaces de:

- Identificar y describir las etapas fundamentales del proceso de minería de datos, incluyendo la recopilación, limpieza, transformación y análisis de datos.
- Identificar y modelar las diferentes tareas de minería de datos (e.g., tareas descriptivas, predictivas, etc.).
- Aplicar técnicas y modelos de aprendizaje automático de distinto tipo.
- Evaluar y comparar diferentes técnicas y modelos de aprendizaje automático.

III. HABILIDADES RECOMENDADAS

- Conocimientos básicos de estadística.
- Nociones básicas de programación en Python.
- Conceptos fundamentales de bases de datos.
- Nivel intermedio-avanzado de lectura en inglés.

IV. METODOLOGÍA Y ESTRATEGIA DE EVALUACIÓN

El curso contempla una sesión semanal de 2 horas y 15 minutos, que incluye un breve descanso intermedio. Durante esta sesión, se pueden programar diferentes actividades, como clases expositivas, talleres prácticos o tutorías personalizadas.

El curso contempla una estrategia de evaluación diversa e integral. Todas las evaluaciones son **individuales**. En primer lugar, se realizará un estudio de caso que representa el 20% de la evaluación del curso. Luego, se llevará a cabo un desafío de predicción que funciona como un taller práctico de minería de datos que representa el 30% de la evaluación final. El desafío del año pasado se puede revisar en GitHub (https://github.com/Doctorado-UDP/data-mining-2024/tree/main/prediction_challenge).

Finalmente, los estudiantes trabajarán en mini-proyectos de minería de datos que les permitirán aplicar conceptos aprendidos a problemas reales. Cada estudiante definirá una pregunta de investigación original y desarrollará un proyecto que la aborde, lo que implicará recopilar y analizar datos relevantes, establecer objetivos y métricas para evaluar el éxito del proyecto y diseñar las etapas necesarias para alcanzarlos. Los mini-proyectos serán evaluados según lo siguiente: informe de resultados (20%), presentación en clase (10%), coevaluación realizada por pares (10%) y autoevaluación del estudiante (10%).

Para cada entrega, se recomienda, cuando sea posible, subir al GitHub del curso (https://github.com/Doctorado-UDP/data-mining-2025) lo siguiente: bases de datos utilizadas, códigos en Python y/o R e informes en LaTeX y/o Markdown.

Esta estrategia de evaluación permitirá proporcionar una retroalimentación constructiva y personalizada a cada estudiante, lo que ayudará a mejorar su comprensión y habilidades en minería de datos.

V. CRONOGRAMA DE TRABAJO

Sesión 1. Introducción al curso y minería de datos
► Jueves 21 de agosto de 2025

La clase comenzará con una breve introducción sobre el curso y la minería de datos, su relación con el aprendizaje automático y su relevancia en el mundo actual. Además, esta sesión presentará las diferentes etapas del proceso de minería de datos e investigación: definición del problema, recopilación de datos, procesamiento, modelado, evaluación y despliegue, relacionándolas con áreas de la investigación social.

Sesión 2. Revisión de control de versiones con Git y herramientas de producción de contenido con LaTeX y Markdown
► Jueves 28 de agosto de 2025

En esta sesión se introducirán las herramientas relevantes para los procesos de minería y análisis de datos, incluyendo Git como un sistema de control de versiones, y LaTeX y Markdown como herramientas de producción de contenido.

Sesión 3. Análisis exploratorio de datos
► Jueves 4 de septiembre de 2025

Esta sesión se centrará en el análisis exploratorio de datos, abarcando temas como la descripción y manipulación de diferentes tipos de datos, técnicas de limpieza y procesamiento, imputación de valores faltantes, detección y tratamiento de *outliers*, entre otros.

Sesión 4. Modelamiento descriptivo
► Jueves 11 de septiembre de 2025

Esta sesión se centrará en tareas de minería de datos que se enfocan en el análisis descriptivo, incluyendo técnicas no supervisadas de aprendizaje automático y *clustering*. Se explorarán diferentes enfoques para entender la estructura y patrones en los datos

Sesión 5. Presentación de estudio de caso y análisis de ejemplos de minería de datos exitosos
► Jueves 25 de septiembre de 2025

En esta sesión, los estudiantes seleccionarán un caso real donde se haya aplicado con éxito la minería de datos y deberán preparar un estudio de caso detallado. El análisis incluirá el problema identificado, las etapas del proceso de minería de datos que se utilizaron, las técnicas y herramientas empleadas y los resultados obtenidos.

Cada estudiante presentará su estudio de caso durante la sesión, seguido de una reflexión crítica sobre posibles mejoras en la aplicación y otros enfoques que podrían haber sido utilizados (**20% de la nota final del curso**). La presentación debe ser de 15 minutos. Esta discusión generará interés y conexión entre la teoría y la práctica, permitiendo a los estudiantes analizar casos reales y aprender de las experiencias de sus compañeros.

Sesión 6. Introducción a los algoritmos de aprendizaje automático supervisado
► Clase recuperativa, fecha por determinar

En esta sesión, se llevará a cabo una introducción exhaustiva y detallada del aprendizaje automático supervisado. Se explorarán los conceptos fundamentales de este tipo de machine learning, incluyendo la definición de problemas de clasificación y regresión, la selección de parámetros y la preparación de los datos para el entrenamiento de los modelos.

Sesión 7. Tareas de clasificación con algoritmos y modelos de aprendizaje automático I
► Jueves 2 de octubre de 2025

En esta sesión, los estudiantes aprenderán sobre el modelado predictivo en tareas de clasificación utilizando algoritmos clasificadores. Se explorarán diferentes aproximaciones para resolver problemas de clasificación, incluyendo regresiones logísticas y naïve Bayes. Los estudiantes también aprenderán cómo elegir entre estos algoritmos según las características del problema y los datos disponibles.

Sesión 8. Tareas de clasificación con algoritmos y modelos de aprendizaje automático II
► Jueves 9 de octubre de 2025

En esta sesión, los estudiantes aprenderán sobre el modelamiento predictivo en tareas de clasificación utilizando algoritmos clasificadores más complejos. Se explorarán aproximaciones para resolver problemas de clasificación, incluyendo árboles de decisión, Support Vector Machine (SVM) y redes neuronales como Extreme Gradient Boosting (XGBoost). Los estudiantes también aprenderán cómo elegir entre estos algoritmos según las características del problema y los datos disponibles.

Sesión 9. Métricas de evaluación de modelos de aprendizaje automático
► Jueves 16 de octubre de 2025

En esta sesión, se presentará un panorama amplio de las métricas de evaluación de modelos utilizadas en aprendizaje automático. Se explorarán diferentes métricas para evaluar el rendimiento de los modelos, incluyendo precisión, recall y F1-score, así como otros métodos de evaluación más complejos como *k-fold cross-validation*.

Además, se analizarán ejemplos prácticos de cómo utilizar estas métricas para comparar el desempeño de diferentes modelos en problemas reales. El objetivo es proporcionar una comprensión profunda de las técnicas de evaluación y su aplicación efectiva en minería de datos.

Sesión 10. Ensamblado de modelos y técnicas avanzadas de predicción
► Jueves 23 de octubre de 2025

En esta sesión, se explorarán las estrategias para ensamblar modelos y mejorar su rendimiento. Los estudiantes aprenderán cómo combinar diferentes modelos para crear una predicción más robusta. Además, se presentarán técnicas avanzadas de predicción como las ventanas móviles (*rolling windows*) y el ajuste de parámetros (*hyperparameter tuning*), que permiten adaptar los modelos a diferentes contextos y problemas.

También se analizarán ejemplos prácticos de cómo ensamblar modelos utilizando técnicas como *bagging, boosting* y *stacking*. Se discutirán las ventajas y desventajas de cada enfoque y se presentará un panorama general sobre la selección del más adecuado para diferentes problemas de minería de datos.

Sesión 11. Desafío de predicción
► Jueves 30 de octubre de 2025

En esta sesión, se presentará el desafío de predicción que opera como un taller práctico que permitirá aplicar las habilidades adquiridas durante el curso a un problema real. El desafío del año pasado se puede revisar en GitHub (https://github.com/Doctorado-UDP/data-mining-2024/tree/main/prediction_challenge).

Los estudiantes deberán utilizar el conjunto de datos que se revelará para entrenar y validar sus propios modelos de aprendizaje automático en Python. Se presentará un script base que los estudiantes pueden refinar y usar para entrenar y preparar sus propios modelos supervisados. Diariamente se actualizará la tabla de clasificación (*leaderboard*) incluyendo los nuevos modelos preparados por los estudiantes.

Los estudiantes deben entrenar al menos un modelo durante el desafío y pueden realizar hasta un máximo de un envío por día. Otras cuestiones logísticas relacionadas con los datos (e.g., límite de predictores, tipos de recodificación, etc.) serán comunicadas durante esta sesión.

El desafío representa un **30% de la evaluación final del curso** y la nota será calculada con base en el mejor F1-score obtenido por cada estudiante considerando todos los envíos realizados a través de GitHub. El desafío tendrá una duración de **dos semanas**.

Sesión 12. Text-as-Data y NLP para análisis de datos
► Jueves 6 de noviembre de 2025

En esta sesión, exploraremos aproximaciones para analizar grandes volúmenes de texto como si fueran conjuntos de datos numéricos. Descubriremos cómo utilizar técnicas de procesamiento de lenguaje natural (NLP por sus siglas en inglés) para convertir textos en formatos manejables, y luego aprenderemos a aplicar algoritmos de minería de datos a estos conjuntos textuales. También exploraremos herramientas y librerías para automatizar el procesamiento del lenguaje natural y mejorar la extracción de información valiosa de textos en Python.

Sesión 13. Modelos semi-supervisados y clasificación con LLMs
► Jueves 13 de noviembre de 2025

En esta sesión, profundizaremos, aunque de forma introductoria, en los modelos semi-supervisados y su aplicación en la clasificación de textos. Aprenderemos a utilizar transformadores como BERT y RoBERTa y grandes modelos lenguaje (LLMs por sus siglas en inglés) para mejorar el rendimiento de algoritmos de clasificación en problemas de minería de datos. Luego aplicaremos estos conceptos en ejercicios prácticos.

Sesión 14. Mini-Proyectos de aplicación en minería de datos
► Jueves 20 de noviembre de 2025

En esta sesión, los estudiantes comenzarán a trabajar en mini-proyectos de minería de datos que les permitan aplicar conceptos aprendidos durante el curso a problemas reales. Cada estudiante definirá una pregunta de investigación original y desarrollará un proyecto que aborde

esta pregunta. Esto implicará recopilar y analizar datos relevantes, establecer objetivos y métricas para evaluar el éxito del proyecto, y diseñar las etapas necesarias para alcanzarlos.

Durante la sesión, los estudiantes recibirán retroalimentación constructiva sobre sus proyectos en curso, lo que les permitirá ajustar su enfoque y mejorar su plan de trabajo.

Sesión 15. Desarrollo del proyecto de minería de datos
► Jueves 27 de noviembre de 2025

En esta sesión, los estudiantes continuarán trabajando en sus mini-proyectos, aplicando las habilidades y herramientas aprendidas durante el curso para desarrollar su trabajo. El objetivo es llevar a cabo la preparación de datos, el modelado y la exploración inicial de los resultados.

Durante la sesión, los estudiantes recibirán retroalimentación constructiva sobre su progreso. Se revisarán algunos de los obstáculos enfrentados y se sugerirán ajustes que mejoren el proceso para todos los estudiantes. Al finalizar la sesión, cada estudiante deberá mostrar el progreso logrado hasta el momento y recibir retroalimentación valiosa sobre su trabajo en curso.

Sesión 16. Presentación de proyectos de minería de datos
► Jueves 4 de diciembre de 2025

Antes de la sesión, los estudiantes deben entregar un informe detallado sobre los resultados de su mini-proyecto y respaldar el código y datos en GitHub si corresponde. Esto representa el **20% de la evaluación final del curso**. Durante la sesión, cada estudiante presentará su proyecto a la clase, demostrando su proceso de análisis, resultados y conclusiones. Cada presentación tendrá un límite de tiempo de 20 minutos, seguida de una sesión de preguntas y respuestas donde el resto de la clase podrá plantear dudas o comentarios, lo que representa el **10% de la evaluación final del curso**.

Además, se incentivará a los estudiantes para incluir consideraciones sobre el impacto en el campo de estudio y su relevancia práctica en el contexto real. Se les pedirá a los estudiantes que completen una coevaluación de sus pares (**10% de la evaluación final del curso**) y personal sobre su proyecto (**autoevaluación, 10% de la evaluación final del curso**).

Sesión 17. Cierre y análisis retrospectivo del curso
► Jueves 11 de diciembre de 2025

En esta sesión, los estudiantes tendrán la oportunidad de recapitular todos los conceptos aprendidos durante las sesiones anteriores. Se realizará un análisis retrospectivo del curso, revisando los conceptos clave aprendidos y reflexionando sobre cómo pueden ser aplicados en el contexto académico o profesional. Esta sesión finalizará con una visión general de las posibilidades y oportunidades que ofrece la minería de datos en diferentes campos y sectores, preparando a los estudiantes para su integración laboral o perfeccionamiento académico.

VI. BIBLIOGRAFÍA.

- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis, 31*(3):337-351.
- Barrie, C., Palaiologou, E., & Törnberg, P. (2024). *Prompt Stability Scoring for Text Annotation with Large Language Models*. arXiv preprint, Cornell University.
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30):e2305016120.
- González-Bustamante, B. (2024). *Benchmarking LLMs in Political Content Text-Annotation: Proof-of-Concept with Toxicity and Incivility Data*. arXiv preprint, Cornell University.

- González-Bustamante, B. (2025). Machine Learning and Political Events: Application of a Semi-supervised Approach to Produce a Dataset on Presidential Cabinets. *Social Science Computer Review*. OnlineFirst.
- Grimmer J., & Stewart B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267-297.
- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as Data: A News Framework for Machine Learning and the Social Sciences*. Princeton University Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd Edition*. Springer.
- Hvitfeldt, E., & Silge, J. (2022). *Supervised Machine Learning for Text Analysis in R*. CRC Press.
- Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature, 616*(7957): 413-413.
- Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining, 2nd edition*. Pearson.
- Timoneda, J. C., & Vallejo Vera, S. (2024). BERT, RoBERTa or DeBERTa? Comparing Performance Across Transformer Models in Political Science Text. *The Journal of Politics*. OnlineFirst.
- Törnberg, P. (2023). *ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning*. arXiv preprint, Cornell University.

VII. PROFESOR Y HORARIO DE OFICINA

Bastián González-Bustamante. Investigador postdoctoral en Ciencias Sociales Computacionales y profesor de Gobernanza y Desarrollo en el Instituto de Administración Pública de la Facultad de Gobernanza y Asuntos Globales de la Universidad de Leiden, Países Bajos. Además, en Chile, se desempaña como académico en la Facultad de Administración y Economía de la Universidad Diego Portales (UDP). Doctor (DPhil/PhD) en Ciencia Política por la Universidad de Oxford, Reino Unido. Magíster en Ciencia Política, Administrador Público y Licenciado en Ciencias Políticas y Gubernamentales por la Universidad de Chile.

Página Web: https://bgonzalezbustamante.com
Email: bastian.gonzalez.b@mail.udp.cl

Si desea programar una reunión personal, puede hacerlo con un plazo de hasta **24 horas de anticipación**, siempre que exista disponibilidad:

- https://app.usemotion.com/meet/bgonzalezbustamante/meeting