



## DOA0006 – Minería de Datos


### Presentación



 Programa: Doctorado en Ciencias de la Administración, Facultad de Administración y Economía, Universidad Diego Portales (UDP).

 Profesor: [Bastían González-Bustamante](#).

 Semestre: Privavera 2024.

 Ubicación: *Online* la mayor parte del semestre.

 Día y horario: Jueves de 16.00 a 18.15 hrs. (GMT-04.00, Santiago, Chile).

 [Descargar última versión](#)  última actualización 26 de julio de 2024).

### Descripción

En este curso de minería de datos, a lo largo de 16 sesiones, 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 procesamiento de datos, 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 del 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 cantidades de datos y estarán capacitados para aplicar estos conocimientos en distintos contextos.

## Objetivos de aprendizaje

- 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 minería de datos.

## Requisitos recomendados

- Conocimientos básicos de estadística.
- Nociones básicas de programación en Python y R.
- Conceptos fundamentales de bases de datos.
- Nivel intermedio-avanzado de inglés en habilidad lectora.

## Metodología y estrategia de evaluación

El curso contempla una sesión semanal de ⌚ 2h 📺 15m, que incluye un breve descanso (☞ *break*) intermedio. Durante esta sesión, se pueden programar diferentes actividades, como clases expositivas, talleres prácticos o tutorías personalizadas. Además, algunas sesiones contarán con invitados que expondrán sobre sus áreas de investigación.

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 representará el 20 % de la evaluación del curso. Luego, se llevará a cabo un taller práctico de minería de datos que valdrá el 30 % de la evaluación del curso.

Finalmente, los estudiantes trabajarán en mini-proyectos de minería de datos que les permitan 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 %). En la sección **Calendario y contenidos** del programa, encontrará marcadas las fechas y detalles específicos para cada evaluación con el símbolo 📅.

Para cada entrega, es necesario subir al 🌐 [GitHub del curso](#) lo siguiente: bases de datos utilizadas, códigos en Python y/o R e informes en  $\text{\LaTeX}$  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.

## Calendario y contenidos

### Sesión 1. Introducción al curso y minería de datos

📅 Jueves 22 de agosto

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 $\text{\LaTeX}$ y Markdown

📅 Jueves 29 de agosto

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  $\text{\LaTeX}$  y Markdown como herramientas de producción de contenido.

### Sesión 3. Análisis exploratorio de datos

📅 Jueves 5 de septiembre

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 12 de septiembre

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

📅 Viernes 27 de septiembre (🔴 por confirmar)

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, seguida de una reflexión crítica sobre posibles mejoras en la aplicación y otros enfoques que podrían haber sido utilizados (**20 % del valor final del curso**). La presentación debe ser de ⌚ 15m. 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

📅 Jueves 3 de octubre

En esta sesión, se llevará a cabo una introducción exhaustiva y detallada al 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 características y la preparación de los datos para el entrenamiento de modelos.

## Sesión 7. Tareas de clasificación con algoritmos probabilísticos y modelamiento predictivo

📅 Viernes 4 de octubre (🚫 por confirmar)

En esta sesión, los estudiantes aprenderán sobre el modelado predictivo en tareas de clasificación utilizando algoritmos clasificadores probabilísticos. 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 discriminativos y modelamiento predictivo

📅 Jueves 10 de octubre

En esta sesión, los estudiantes aprenderán sobre el modelamiento predictivo en tareas de clasificación utilizando algoritmos clasificadores discriminativos. Se explorarán diferentes aproximaciones para resolver problemas de clasificación, incluyendo árboles de decisión, Support Vector Machine (SVM) y redes neurales 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. Técnicas de evaluación de modelos en minería de datos

📅 Jueves 17 de octubre

En esta sesión, se presentará un panorama amplio de las técnicas de evaluación de modelos utilizadas en minería de datos. Se explorarán diferentes métricas para evaluar el rendimiento de los modelos, incluyendo la precisión, el recall y el 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 24 de octubre

En esta sesión, se explorarán las estrategias para ensamblar modelos y mejorar su rendimiento. Los estudiantes aprenderán cómo combinar diferentes modelos para crear un modelo más poderoso y robusto. 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.

Además, 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 mejor enfoque para diferentes problemas de minería de datos.

### Sesión 11. Taller práctico de minería de datos

📅 Jueves 7 de noviembre

En esta sesión, los estudiantes participarán en un taller práctico integral que les permitirá aplicar los conceptos y habilidades aprendidos durante el curso a un problema real. Los estudiantes utilizarán una herramienta de minería de datos elegida previamente (Python y/o R) para analizar un conjunto de datos proporcionado.

El objetivo es que los estudiantes puedan mostrar su comprensión práctica del proceso de minería de datos, desde la selección de la herramienta y el análisis de los datos hasta la presentación de los resultados. La sesión será evaluada con un **30 % de la calificación final del curso**.

Durante el taller, los estudiantes recibirán retroalimentación constructiva de sus compañeros y del profesor sobre su proceso de trabajo, lo que les permitirá mejorar y refinar sus habilidades en minería de datos. Al finalizar la sesión, cada estudiante tendrá 📅 **una semana** para presentar, a través de [GitHub](#), un informe detallado de su análisis, incluyendo resultados obtenidos y las conclusiones alcanzadas.

### Sesión 12. Text-as-Data y NLP para análisis de datos

📅 Jueves 14 de noviembre

En esta sesión, exploraremos las habilidades para analizar grandes cantidades de texto como si fuera un conjunto de datos numéricos. Descubriremos cómo utilizar técnicas de procesamiento del lenguaje natural (NLP) para convertir textos en formatos manejables, y luego aprenderemos a aplicar algoritmos de minería de datos a estos conjuntos de texto. También exploraremos herramientas y bibliotecas como NLTK, spaCy y gensim para automatizar el procesamiento del lenguaje natural y mejorar la extracción de información valiosa de textos grandes en Python.

### Sesión 13. Modelos semi-supervisados y clasificación con LLMs

📅 Jueves 21 de noviembre

En esta sesión, profundizaremos en los modelos semi-supervisados y su aplicación en la clasificación de textos. Aprenderemos a utilizar transformadores como BERT y RoBERTa y grandes modelos de lenguaje (LLMs) 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 28 de noviembre

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 esa 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 en minería de datos



📅 Jueves 5 de diciembre

En esta sesión, los estudiantes continuarán trabajando en sus mini-proyectos, aplicando las habilidades y herramientas aprendidas a lo largo del 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, tanto del profesor como de sus compañeros. 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 ese momento y recibir retroalimentación valiosa sobre su trabajo en curso.

## Sesión 16. Presentación de proyectos en minería de datos

📅 Jueves 12 de diciembre

Antes de la sesión, los estudiantes deben entregar un informe detallado, a través de  [GitHub](#), sobre los resultados de su mini-proyecto, lo que 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  20m, seguido 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, facilitando un espacio para el enriquecimiento entre pares. 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**).

Esta sesión proporcionará un espacio para compartir conocimientos y experiencias entre pares, lo que ayudará a profundizar en el aprendizaje y a mejorar las habilidades de los estudiantes.

## Bibliografía


Grimmer, J., Roberets, M. E., & Stewart, B. M. (2022). *Text as Data: A New 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.

Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining, 2nd edition*. Pearson.

## Integridad académica

 Deberá cumplir íntegramente con las normas de integridad académica relativas a la colaboración y el plagio, siguiendo las normas establecidas por la Facultad de Administración y Economía y la Universidad Diego Portales (UDP).

## Horario de oficina

■ Si desea programar una reunión personal, puede hacerlo con un plazo de hasta **48 horas de anticipación**, siempre que exista disponibilidad:

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

## Sobre el profesor

### Bastían González-Bustamante

🏠 [bgonzalezbustamante.com](https://bgonzalezbustamante.com) ✉ [bastian.gonzalez.b@mail.udp.cl](mailto:bastian.gonzalez.b@mail.udp.cl)  
🗣 [@bgonzalezbustamante](#) 📦 [OSF](#) 📄 [Google Scholar](#)

Investigador posdoctoral 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. 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.

Además, en Chile, se desempeña como académico en la Facultad de Administración y Economía de la Universidad Diego Portales (UDP) e investigador asociado en el [Training Data Lab](#), un grupo de investigación centrado en la minería de datos, la modelización econométrica y el aprendizaje automático e inteligencia artificial en ciencias sociales.

Sus intereses de investigación se encuentran en la intersección de la política comparada y el gobierno, centrándose principalmente en los gabinetes, los regímenes políticos y los servicios civiles. Metodológicamente, sus intereses se basan en la aplicación del análisis cuantitativo de textos, los métodos de aprendizaje automático y las estrategias de inferencia causal en el campo de la política comparada. En los últimos años, su trabajo ha sido publicado en *The International Journal of Press/Politics*, *World Development*, *Government and Opposition*, *The British Journal of Politics and International Relations*, *Bulletin of Latin American Research* y otras revistas.

---

© Este programa de curso está bajo una licencia [Creative Commons Atribución-NoComercial 4.0 Internacional](#). Fue creado a partir del programa original del curso del [Doctorado en Ciencias de la Administración de la UDP](#), con la colaboración de edutekaLab y el [modelo GPT-3.5 de OpenAI](#). Posteriormente, fue editado por [B González-Bustamante](#) con la ayuda del modelo [Llama-3](#).