Euncet Business School

centro adscrito a:

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

GUIA DOCENTE DE Minería de Datos 2026-27

DATOS GENERALES

Número:	Minería de Datos
Codigo:	801363 (BUSINESSTECH)
Curso:	2026-27
Titulación:	Grado en Empresa, Innovación y Tecnología
Nº de créditos (ECTS):	6
Ubicación en el plan de estudios:	3º. Curso, 2º. cuatrimestre
Departamento:	Métodos Cuantitativos
Responsable departamento:	Dr. Iván Romero
Fecha de la última revisión:	Julio 2026
Profesor Responsable:	Xavi Royo Melero

1. DESCRIPCIÓN GENERAL

La minería de datos constituye una de las disciplinas centrales de la analítica de negocio, al integrar técnicas estadísticas, computacionales y de aprendizaje automático para extraer conocimiento accionable a partir de grandes volúmenes de datos. En un entorno empresarial donde la generación de datos crece de forma exponencial, la capacidad de descubrir patrones, predecir comportamientos y segmentar colectivos resulta indispensable para la toma de decisiones estratégicas.

La asignatura se estructura en seis bloques temáticos que recorren el ciclo completo de un proyecto de minería de datos: desde la comprensión del problema y los fundamentos del machine learning, pasando por el preprocesamiento y la exploración de los datos (EDA), hasta la construcción y evaluación de modelos predictivos y descriptivos, así como el descubrimiento de patrones mediante técnicas de segmentación, asociación y recomendación. A lo largo del curso se empleará el lenguaje Python como herramienta principal, con librerías como pandas, scikit-learn, matplotlib y seaborn.

El primer bloque proporciona el marco conceptual de la minería de datos y el machine learning, diferenciando los tipos de aprendizaje supervisado y no supervisado, y situando la disciplina dentro del ecosistema más amplio de la ciencia de datos.

El segundo bloque aborda el preprocesamiento y la exploración de datos, fase crítica que determina la calidad y relevancia de cualquier modelo posterior. Se estudian técnicas de limpieza, transformación, tratamiento de valores ausentes y visualización exploratoria.

Los bloques tercero y cuarto se centran en los modelos supervisados: clasificación y regresión, respectivamente. Se presentan los algoritmos más relevantes (regresión logística y lineal, árboles de decisión y random forest), métricas de evaluación y criterios de selección según el problema de negocio.

El quinto bloque introduce las técnicas de aprendizaje no supervisado, con especial énfasis en el clustering (K-Means y clustering jerárquico) y la segmentación de clientes, herramienta clave para el análisis de negocio.

El sexto bloque aborda técnicas de descubrimiento de patrones y personalización basadas en reglas de asociación y sistemas de recomendación. Se estudia el algoritmo Apriori para el análisis de relaciones entre productos y los enfoques de collaborative filtering y content-based filtering para la generación de recomendaciones, examinando su utilidad en distintos contextos de negocio.

2. OBJETIVOS

Al finalizar el curso el estudiante será capaz de:

- Comprender los fundamentos conceptuales del machine learning y la minería de datos, diferenciando tipos de aprendizaje y paradigmas algorítmicos.
- Aplicar técnicas de preprocesamiento y exploración de datos para garantizar la calidad e idoneidad de los datasets antes del modelado.
- Construir, entrenar y evaluar modelos de clasificación y regresión utilizando Python y la librería scikit-learn.
- Aplicar algoritmos de clustering y segmentación para identificar grupos naturales en los datos con orientación al análisis de negocio.
- Aplicar reglas de asociación y sistemas de recomendación para descubrir patrones de comportamiento y apoyar decisiones de personalización, recomendación y cross-selling.
- Seleccionar el algoritmo y las métricas de evaluación adecuados en función del problema de negocio y las características de los datos.
- Comunicar de forma efectiva los resultados de un análisis de minería de datos, tanto de forma oral como escrita, a audiencias técnicas y no técnicas.
- Desarrollar una visión crítica sobre las limitaciones, sesgos y consideraciones éticas asociadas a la aplicación de modelos predictivos en entornos empresariales.

3. RESULTADOS DEL APRENDIZAJE

- Identificar y diferenciar los principales tipos de problemas de minería de datos (clasificación, regresión, clustering y asociación) y los paradigmas de machine learning asociados.
- Aplicar un proceso estructurado de preprocesamiento de datos (limpieza, codificación, escalado y partición de datos) previo a la construcción de modelos.
- Construir pipelines de modelado supervisado, evaluar su rendimiento mediante métricas apropiadas y aplicar técnicas de validación cruzada y ajuste de hiperparámetros.
- Implementar técnicas de segmentación y clustering, interpretar los grupos resultantes y trasladar los hallazgos a recomendaciones de negocio.
- Aplicar reglas de asociación y sistemas de recomendación básicos, interpretando sus resultados en contextos de negocio.
- Reflexionar críticamente sobre los resultados de los modelos, identificando posibles sesgos, limitaciones y requisitos éticos en su aplicación empresarial.

4. CONTENIDOS

TEMA 1. INTRODUCCIÓN A LA MINERÍA DE DATOS Y MACHINE LEARNING

Resultados específicos del aprendizaje

El estudiante después de estudiar el tema y realizar los ejercicios, será capaz de:

- Definir los conceptos clave de minería de datos, machine learning e inteligencia artificial, situándolos en el ecosistema de la ciencia de datos.
- Distinguir entre aprendizaje supervisado y no supervisado, identificando casos de uso empresarial para cada paradigma.
- Describir las fases de un proyecto de minería de datos (CRISP-DM) y relacionarlas con el ciclo de vida de un modelo en producción.
- Identificar las principales librerías de Python para minería de datos (pandas, scikit-learn, matplotlib, seaborn) y su rol en el flujo de trabajo.

Contenido

1.1. Conceptos fundamentales: datos, información, conocimiento, minería de datos e IA.
1.2. Tipos de aprendizaje: supervisado y no supervisado.
1.3. Taxonomía de problemas: clasificación, regresión, clustering y asociación.

1.4. Metodología CRISP-DM y ciclo de vida de un proyecto de datos.
1.5. Despliegue de modelos como fase final del ciclo de vida: concepto y visión general.
1.6. Ecosistema Python para minería de datos: pandas, scikit-learn, matplotlib y seaborn.

TEMA 2. PREPROCESAMIENTO Y EXPLORACIÓN DE DATOS (EDA)

Resultados específicos del aprendizaje

El estudiante después de estudiar el tema y realizar los ejercicios, será capaz de:

- Realizar un análisis exploratorio de datos (EDA) completo, identificando la distribución, relaciones y anomalías presentes en un dataset.
- Aplicar técnicas de limpieza de datos: detección y tratamiento de valores ausentes, duplicados y outliers.
- Transformar variables mediante codificación (one-hot, label encoding) y escalado (normalización, estandarización).
- Utilizar visualizaciones estadísticas (histogramas, boxplots, heatmaps de correlación) para comunicar hallazgos del EDA.

Contenido

2.1. Análisis exploratorio de datos (EDA): estadísticas descriptivas y visualización.
2.2. Detección y tratamiento de valores ausentes y duplicados.
2.3. Identificación y gestión de outliers.
2.4. Codificación de variables categóricas: one-hot encoding y label encoding.
2.5. Escalado y normalización de variables: MinMaxScaler y StandardScaler.
2.6. División de datos: train/test split y prevención de data leakage.
2.7. Pipelines de preprocesamiento con scikit-learn: diseño y aplicación.

TEMA 3. MODELOS DE CLASIFICACIÓN

Resultados específicos del aprendizaje

El estudiante después de estudiar el tema y realizar los ejercicios, será capaz de:

- Comprender el funcionamiento de los principales algoritmos de clasificación: regresión logística, árboles de decisión y random forest.
- Entrenar y evaluar modelos de clasificación utilizando scikit-learn, aplicando validación cruzada.
- Interpretar métricas de evaluación de clasificación: accuracy, precision, recall, F1-score, AUC-ROC y matriz de confusión.

- Seleccionar el algoritmo de clasificación más adecuado en función de las características del problema de negocio y del dataset.

Contenido

3.1. Fundamentos de la clasificación supervisada y aplicaciones empresariales: detección de fraude, concesión de créditos y predicción de abandono de clientes.
3.2. Regresión logística: fundamentos e interpretación.
3.3. Árboles de decisión: criterios de división y control del overfitting.
3.4. Random forest como ensemble method.
3.5. Métricas de evaluación: accuracy, precision, recall, F1-score, AUC-ROC y matriz de confusión.
3.6. Validación cruzada, selección de modelos, ajuste de hiperparámetros.

TEMA 4. MODELOS DE REGRESIÓN

Resultados específicos del aprendizaje

El estudiante después de estudiar el tema y realizar los ejercicios, será capaz de:

- Comprender los fundamentos de la regresión lineal simple y múltiple.
- Aplicar técnicas de regularización para gestionar el sobreajuste: Ridge y Lasso
- Entrenar y evaluar modelos de regresión utilizando scikit-learn: regresión lineal, árboles de regresión y métodos ensemble.
- Interpretar métricas de evaluación de regresión: MAE, RMSE y R^2.
- Seleccionar el algoritmo más adecuado según el problema de negocio y las características de los datos.
- Comprender e interpretar los resultados de un modelo de regresión para la toma de decisiones.

Contenido

4.1. Regresión lineal simple y múltiple: fundamentos e interpretación.
4.2. Regresión regularizada: Ridge y Lasso como técnicas para evitar el sobreajuste.
4.3. Árboles de regresión y métodos ensemble para regresión.
4.4. Métricas de evaluación: MAE, RMSE, R^2.
4.5. Criterios de selección del algoritmo según el problema de negocio.
4.6. Comprensión de modelos de regresión: importancia de variables y comunicación de resultados.

TEMA 5. TÉCNICAS DE CLUSTERING Y SEGMENTACIÓN

Resultados específicos del aprendizaje

El estudiante después de estudiar el tema y realizar los ejercicios, será capaz de:

- Comprender los fundamentos del aprendizaje no supervisado y distinguirlo del aprendizaje supervisado.
- Aplicar el algoritmo K-Means para segmentar datasets, determinando el número óptimo de clusters mediante el método del codo y el índice de silueta.
- Implementar clustering jerárquico e interpretar dendrogramas para la toma de decisiones sobre el número de grupos.
- Comprender el concepto de reducción de dimensionalidad mediante PCA y su utilidad como paso previo al clustering.
- Trasladar los resultados del clustering a perfiles de segmentos accionables para el negocio.

Contenido

5.1. Fundamentos del aprendizaje no supervisado y aplicaciones de negocio.
5.2. K-Means: algoritmo, selección de K mediante el método del codo e índice de silueta
5.3. Clustering jerárquico aglomerativo y dendrogramas.
5.4. Reducción de dimensionalidad mediante PCA: concepto y aplicación en clustering.
5.5. Segmentación de clientes: RFM (Recency, Frequency, Monetary) y aplicaciones de marketing.
5.6. Interpretación y comunicación de perfiles de segmentos para la toma de decisiones.

TEMA 6. REGLAS DE ASOCIACIÓN Y SISTEMAS DE RECOMENDACIÓN

Resultados específicos del aprendizaje

El estudiante después de estudiar el tema y realizar los ejercicios, será capaz de:

- Comprender el concepto de regla de asociación e identificar aplicaciones empresariales relevantes.
- Aplicar el algoritmo Apriori para descubrir patrones de comportamiento en datos de transacciones.
- Interpretar las métricas clave de las reglas de asociación: soporte, confianza y lift.
- Distinguir entre collaborative filtering y content-based filtering, identificando casos de uso para cada enfoque.
- Trasladar los resultados de un sistema de recomendación a decisiones de negocio: recomendación de productos, personalización y cross-selling.

Contenido

6.1. Aprendizaje no supervisado basado en asociación: descubrimiento de patrones y aplicaciones empresariales.
6.2. Reglas de asociación y análisis de cesta de la compra: algoritmo Apriori y métricas clave.
6.3. Sistemas de recomendación: collaborative filtering y content-based filtering.
6.4. Aplicaciones empresariales: recomendación de productos, personalización y cross-selling.
6.5. Interpretación y comunicación de resultados para la toma de decisiones de negocio.

5. METODOLOGÍA DE ENSEÑANZA Y APRENDIZAJE

La asignatura Minería de Datos, situada en el tercer curso de la mención de Analítica de Negocio, integra técnicas estadísticas, computacionales y de aprendizaje automático para extraer conocimiento accionable a partir de grandes volúmenes de datos. Dado su carácter eminentemente aplicado, la modalidad semipresencial combina sesiones presenciales de trabajo práctico intensivo con sesiones asíncronas que permiten consolidar el aprendizaje autónomo en Python, siguiendo el ciclo completo de un proyecto de minería de datos (CRISP-DM).

La modalidad se articula en 6 sesiones presenciales en aula física, 6 sesiones asíncronas (o síncronas cuando así lo determine el profesorado) y 1 sesión de examen final presencial, sumando un total de 13 sesiones.

Sesiones presenciales (6 sesiones):

Cada sesión combina una parte expositiva y una parte práctica de un mínimo de hora y media. En la parte teórica, el profesorado introduce los conceptos, algoritmos y marcos metodológicos del bloque temático correspondiente apoyándose en el material docente y los notebooks disponibles en el campus virtual. Se utilizarán casos de uso empresariales reales (detección de fraude, segmentación de clientes, sistemas de recomendación, predicción de churn) para contextualizar cada técnica y facilitar la transferencia al entorno profesional. Se fomentará la participación activa mediante la discusión crítica de resultados, la comparación entre algoritmos y la reflexión sobre las implicaciones éticas y limitaciones de los modelos. En la parte práctica, el estudiante (de forma individual o en grupo) implementará, entrenará y evaluará modelos en Python. El orden de ambas partes podrá variar según criterio pedagógico del profesorado. Al menos 4 de las 6 sesiones presenciales incluirán obligatoriamente parte práctica con actividad evaluable. No se

avisará con antelación del momento exacto en que tendrá lugar la actividad práctica, con el fin de favorecer la asistencia completa a cada sesión.

Sesiones asíncronas (6 sesiones):

Las sesiones asíncronas se articularán mediante materiales docentes, notebooks de Python estructurados por tema y recursos publicados en el campus virtual, preparados específicamente para los contenidos de la asignatura. Al menos 4 de estas 6 sesiones incluirán una parte práctica de un mínimo de hora y media que el estudiante podrá desarrollar de forma asíncrona, o de forma síncrona si así lo decide el profesorado. Al menos 4 sesiones asíncronas contarán con algún tipo de actividad evaluable.

Las principales actividades que se realizarán a lo largo de las sesiones presenciales y asíncronas son:

- Estudio y preparación de los contenidos a partir del material docente y notebooks del campus virtual.
- Clases expositivas (presenciales u asíncronas) sobre fundamentos de machine learning, preprocesamiento, modelado supervisado y no supervisado, y sistemas de recomendación.
- Exploración, limpieza y preprocesamiento de datasets con pandas y scikit-learn (EDA, tratamiento de valores ausentes, codificación, escalado).
- Implementación, entrenamiento y evaluación de modelos supervisados (clasificación y regresión) y no supervisados (clustering, reglas de asociación) en Python.
- Construcción de pipelines completos de modelado con scikit-learn, aplicando validación cruzada y ajuste de hiperparámetros.
- Interpretación y comunicación de resultados técnicos en lenguaje de negocio, adaptados a audiencias no especializadas.
- Entregas prácticas por bloque temático, con elaboración previa autónoma y resolución o presentación obligatoria en sesión.

Se trabajará con notebooks de Python estructurados por tema, disponibles en el campus virtual, para que el estudiante pueda consolidar el aprendizaje autónomo fuera de las sesiones con ejercicios equivalentes a los trabajados en clase. La progresión técnica irá desde el preprocesamiento y los modelos supervisados hasta las técnicas no supervisadas y los sistemas de recomendación.

6. EVALUACIÓN

Las tareas y actividades evaluativas se ajustarán al contenido del material docente expuesto en clase y facilidad en el Campus para comprobar que el alumnado los ha consolidado. De acuerdo con el Plan Bolonia, el modelo premia el esfuerzo constante y continuado del estudiantado. Un 40% de la nota se obtiene de la evaluación continua y el 60% restante, del examen final presencial. El examen final tiene dos convocatorias.

La nota final de la asignatura (NF) se calculará a partir de la siguiente fórmula:

- **NF = Nota Examen Final x 60% + Nota Evaluación Continuada x 40%**
- Nota mínima del examen final para calcular la NF será de 40 puntos sobre 100.
- La asignatura queda aprobada con una NF igual o superior a 50 puntos sobre 100.

Tipo de actividad	Descripción	% Evaluación continua	
Entregas por tema:			**20%**
Ex. Tema 1	Entrega por Classlife	10%	
Ex. Tema 2	Entrega por Classlife	20%	
Ex. Tema 3	Entrega por Classlife	20%	
Ex. Tema 4	Entrega por Classlife	20%	
Ex. Tema 5	Entrega por Classlife	20%	
Ex. Tema 6	Entrega por Classlife	10%	
Examen Parcial I			**10%**
	Examen Parcial I	100%	
Examen Parcial II			**10%**
	Examen Parcial II	100%	
Examen final			**60%**
	Examen final	100%	

7. BIBLIOGRAFÍA

7.1. BIBLIOGRAFÍA BÁSICA

Géron, A. (2025). Hands-On Machine Learning with Scikit-Learn and Pytorch. O'Reilly Media.

Raschka, S. y Mirjalili, V. (2022). Machine Learning with PyTorch and Scikit-Learn. Packt Publishing.

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). Introduction to data mining (2nd ed.). Pearson.