Universidad de Navarra

Machine Learning (Grupo A) (MBDS)

Guía docente 2026-27

PRESENTACIÓN

Breve descripción: En esta asignatura los estudiantes adquirirán destreza en la ejecución de proyectos, tanto individuales como en grupo, donde aplicarán técnicas de Machine Learning (técnicas de aprendizaje supervisado, no supervisado y aprendizaje por refuerzo) a problemas del mundo real. Además, los estudiantes aprenderán técnicas de procesamiento de lenguaje natural (NLP) para manejar y analizar datos textuales, y se introducirá el uso de inteligencia artificial generativa, que permitirá a los estudiantes desarrollar modelos capaces de generar texto, automatizar tareas y utilizar herramientas avanzadas para aplicaciones industriales. En la asignatura se utilizarán los lenguajes R o Python.

- **Titulación**: Máster Universitario en Ciencia de Datos Masivos / Big Data Science
- **Módulo:** Análisis de Datos
- **Materia**: Machine Learning
- **ECTS**: 5
- **Curso, semestre**: curso único del Máster, segundo semestre
- **Carácter**: obligatoria
- **Profesor responsable**: Jesús López Fidalgo
- **Profesorado**: Matías Ávila Clemente
- **Idioma**: castellano
- **Aula, Horario**: ver cronograma

RESULTADOS DE APRENDIZAJE (Competencias)

Competencias Básicas y Generales:

CB6 - Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación

CB7 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio

CB8 - Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios

CB9 - Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades

CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

CG2 - Explorar y explotar tanto tecnológica como estratégicamente de los datos un valor clave para diferentes empresas y organizaciones.

CG4 - Analizar los datos masivos y aportar medidas originales y novedosas para contribuir a la mejora de la competitividad de las empresas e instituciones públicas.

CG5 - Analizar los datos que se generan en el día a día, extrayendo conocimiento de los mismos, realizando predicciones y transformándolos en productos y servicios utilizando las herramientas estadísticas de Data Science.

CG6 - Trabajar en equipos de trabajo o grupos de investigación interdisciplinares de forma eficaz y colaborativa.

CG7 - Conocer y entender las herramientas habituales que se utilizan hoy día en el tratamiento de datos masivos.

CG8 - Saber aplicar los principios éticos relativos a la recogida, almacenamiento, y análisis de datos teniendo en cuenta las posibles discriminaciones directas o indirectas derivadas de la toma de decisiones.

Competencias Específicas

CE3 - Comprender y utilizar algoritmos de Machine Learning en ejemplos prácticos.

CE4.1 - (Estadística) Programar con software estadístico libre R u otro similar y prácticas de cada contenido de esta materia con él.

PROGRAMA

Programa detallado de ML

1. Fundamentos

a. Ciencia de Datos (DS): Contexto del DS, palabras de moda (Buzzwords), roles y flujo de trabajo (Workflow).

b. Comprensión del Negocio: Planteamiento del problema e hipótesis.

c. Comprensión y Preparación de Datos: Conceptos básicos de datos (Data 101), muestreo (Sampling), análisis exploratorio de datos (EDA), preprocesamiento de datos y creación de características (Feature Engineering).

d. Modelado: Introducción a la IA, tipos de aprendizaje automático (ML types of learning), niveles de abstracción en ML y taxonomía de los modelos.

e. Introducción al Aprendizaje por Refuerzo: Comparación entre A/B testing y Multiarmed Bandits.

2._ Algoritmos ML

a. k-Nearest Neighbors (kNN) como clasificador y regresor.

b. Teoría del Aprendizaje Supervisado: algoritmo de ML, descomposición del error, función de coste, riesgos de minimizar, sesgo vs varianza, forma funcional, modelos paramétricos vs no paramétricos (ej. Ridge vs kNN), descomposición del error reducible y entrenamiento.

c. Máxima Verosimilitud y Regresión Lineal: Intuición de la regresión lineal desde una perspectiva diferente. Proyección ortogonal vertical.

d. p<n: Selección de características, proyección de características y reducción de características.

e. Modelos de Regularización: Ridge, Lasso y ElasticNet.

3._ SVM, Naive Bayes y Reglas de Asociación

a. SVM: Clasificador de Margen Máximo, Clasificador de Vectores de Soporte y Máquina de Vectores de Soporte.

b. Naive Bayes: Desde el Teorema de Bayes hasta Naive Bayes.

c. Reglas: Análisis de Asociación y Análisis de Secuencias.

4._ Random Forest y Gradient Boosting

a. Árboles de Decisión: Regresión y clasificación con árboles.

b. Bagging Trees: Conjunto de árboles de decisión. Bootstrap + Aggregation.

c. Random Forest (RF): Mejora sobre Bagging, decorrelación de árboles.

d. Gradient Boosting (GB): Teoría detrás de GB y distintos algoritmos dentro de GB (Extreme GB, Ada Boost...).

5._ Series temporales. Técnicas avanzadas

a. Series Temporales Aplicadas: Historia de las series temporales y modelos, herramientas para el análisis de series temporales, librerías (prophet, skforecast, auto.arima...), entrenamiento de modelos de series temporales, modelado de churn, demanda, etc., usando modelos de ML.

b. Análisis de Supervivencia 101: Introducción, censura, datos truncados, curva de Kaplan-Meier y prueba de Log-Rank.

c. Selección de Características e Interpretabilidad: Importancia de las características, SHAP, y target shuffling.

6._ Modelos no supervisados

a. Teoría de Modelos No Supervisados: Introducción.

b. Distancias: Propiedades de una distancia y ejemplos de distancias.

c. Enfoques básicos de vectorización: Codificación one-hot, Bag of Words (BoW) y TF/IDF.

d. Descomposición Matricial: PCA y SVD (revisión rápida).

e. Análisis Discriminante Lineal (LDA): Clasificador lineal + Proyección ortogonal de características (revisión rápida).

f. Factorización Matricial No Negativa (NMF): Embeddings, NMF y Alternating Least Squares (ALS).

g. Latent Dirichlet Allocation (LDA)

7._ Otros métodos

a. Reducción de la Dimensión No Lineal (Manifold Learning): LLE, ISOMAP, t-SNE y SOMs.

b. Estimación de la Densidad de Probabilidad: Métodos paramétricos, semiparamétricos y no paramétricos.

c. Clustering: k-Means, k-Medoids, Clustering jerárquico, DBSCAN, Affinity Propagation y Silhouette.

8._ Sistemas de recomendación

a. 101 Sistemas de Recomendación: Introducción. Propensity model vs Recommender Systems and examples.

b. Principales tipos de Sistemas de Recomendación: Content-based Recommenders y Collaborative Filtering.

c. Otras formas de recomendar: Latent Factors models y Hybrid Recommender Systems.

d. RecSys usando Deep Learning: CNN, RNN, Autoencoders...

e. Evaluación de Sistemas de Recomendación: Evaluación offline y online, métricas, y pruebas A/B.

9._ Grafos

a. Estudio empírico de Grafos: Introducción, ejemplos y casos reales con grafos. Modelos de ML expresados como grafos.

b. Conceptos básicos de Grafos: Elementos, terminología, tipos de grafos, propiedades y representaciones.

c. Medidas y métricas de Grafos: Métricas básicas de un grafo, centralidad y clustering.

Programa detallado de NLP y Modelos generativos

1_ Técnicas de Preprocesamiento y Representación de Texto

a. Tokenización y diferencias entre los enfoques de tokenización (palabras vs. oraciones)

b. Limpieza y Normalización del Texto: Eliminación de stop words, lematización, y stemming

c. Part of Speech (POS) Tagging y Parsing: Etiquetado de partes del texto y análisis de dependencias

d. Named Entity Recognition (NER): Identificación de entidades nombradas en el texto

e. Representación del Texto (Embeddings avanzados: Word2Vec, GloVe, FastText; Introducción a embeddings basados en Transformers)

f. Análisis de Términos Clave: Técnicas de extracción de keywords para textos

2._ Modelos de Lenguaje y Transfer Learning

a. Introducción a los Modelos de Lenguaje Grandes (LLMs):Arquitectura de Transformers y ejemplos populares: BERT, GPT, Mixtral

b. Transfer Learning: Conceptos y aplicaciones en NLP

c. Fine-tuning de Modelos:Ajuste fino de LLMs con conjuntos de datos específicos

d. Optimización del Entrenamiento: Técnicas de aceleración (precisión mixta y cuantización) ; adaptación de bajo rango (LoRA y QLoRA)

3._ Modelado Temático y Modelos Generativos

a. Topic Modeling: Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), Embeddings + clustering (K-means, HDBSCAN),...

b. Modelos Generativos y Automatización

c. Técnicas Avanzadas de Ingeniería de Prompts: Chain of Thoughts (CoT), Self-Consistency, Tree of Thoughts

d. Uso de Herramientas Externas: Capacidades de los LLMs para llamar funciones externas, Automatización y expansión de las capacidades del modelo

ACTIVIDADES FORMATIVAS

El módulo se compone de clases teóricas y prácticas, además de la realización de un trabajo en grupo.

En las **clases teóricas** el profesor explicarán los contenidos más importantes de la asignatura mediante diapositivas, notebooks, código, referencias a libros y demás recursos que precise la lección.

En las **clases prácticas** los alumnos realizarán las actividades y ejercicios planteados por el profesor.

Habrá un **trabajo en grupo** aplicando las técnicas aprendidas en el módulo.

Actividades formativas	Horas
Clases presenciales teóricas	15
Prácticas con ordenador	20
Trabajos dirigidos	42
Estudio y trabajo personal	40
Tutorías	5
Pruebas presenciales de evaluación	3

Universidad de Navarra

EVALUACIÓN

La evaluación incluirá la asistencia, un proyecto y un examen individual. Se necesita tener **aprobadas individualmente todas** las partes para aprobar la convocatoria ordinaria. Si un alumno copia a otro, suspenderán la asignatura tanto el que copie como el que deja copiarse.

La nota final será un promedio con las siguientes ponderaciones:

EN LA CONVOCATORIA ORDINARIA

- **10%: Asistencia** a clases, seminarios y clases prácticas
- **30%: Proyecto en equipos**. Este se descompone en 60% de nota común para todos los miembros del equipo (20% con la primera entrega y 40% con la segunda entrega) y 40% individual con la defensa. Se evaluará
- **60%: Examen**

EN LA CONVOCATORIA EXTRAORDINARIA

- **10%: Asistencia** a clases, seminarios y clases prácticas
- **30%: Proyecto individual**
- **60%: Examen**

Habrá una actividad voluntaria para subir nota, hasta 0.5 punto, en la nota final de la convocatoria ordinaria (no aplica para la extraordinaria). Será un ejercicio de predicción y según en que decil se ubique sus resultados subirá entre 0.1 y 0.5 punto la nota final.

La matrícula de honor (MH) la recibirán los alumnos con mayor nota final. En caso de empate se realizará un ejercicio de predicción entre los candidatos y aquellos con los mejores resultados obtendrán la MH.

HORARIOS DE ATENCIÓN

Contactar por correo electrónico:

- mavila.3@external.unav.es

BIBLIOGRAFÍA

ML

- James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert, 2014, *An Introduction to Statistical Learning: With Applications in R*, Springer Localízalo en la Biblioteca [Recurso electrónico]. Este libro es una versión simplificada del libro "The Elements of Statistical Learning". Localízalo en la Biblioteca (Recurso electrónico y papel)
- Lantz, Brett, 2015, *Machine Learning with R*, 2nd Edition. Packt Publishing. (Es un libro práctico con muchos ejemplos y con una

explicación sencilla de la teoría.) Localízalo en la Biblioteca. [Recurso electrónico]

- Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* Springer; Edición: 2nd ed. 2009, Localízalo en la Biblioteca [Recurso electrónico] Corr. 9th printing 2017 (19 de mayo de 2017).
- Kevin P. Murphy. Machine Learning A Probabilistic Perspective. MIT Press; Edición: 1 (24 de agosto de 2012). Localízalo en la Biblioteca
- Kuhn, Max & Johnson, Kjell, 2013, *Applied Predictive Modeling*. Springer. (Está en la biblioteca como libro electrónico. Explica la teoría subyacente al paquete de R caret. Es una referencia habitual en estudios de supervised machine learning.) Localízalo en la Biblioteca [Recurso electrónico]

NLP

- Bird, S., Klein E. and Loper E. (2019). Natural Language Processing with Python (1st ed.). https://www.nltk.org/book/ Solutions Book: https://github.com/Sturzgefahr/Natural-Language-Processing-with-Python-Analyzing-Text-with-the-Natural-Language-Toolkit
- Arumugam R., Shanmugmani R. (2018). Hands-On Natural Language Processing with Python: A practical guide to applying deep learning architectures to your NLP applications: Natural Language Processing with Python. https://github.com/PacktPublishing/Hands-On-Natural-Language-Processing-with-Python