

Asignatura

Nombre Asignatura	Minería de Datos II
Código	600522039
Créditos ECTS	6.0

Plan/es donde se imparte

Titulación	Doble Grado en ADE y Análisis de Negocios (Plan 2022)
Carácter	OBLIGATORIA
Curso	3

Datos Generales

➤ PROFESORADO

Gabriel Antonio Valverde Castilla, Doctor en Ingeniería Matemática, Estadística e Investigación Operativa, CDO de Celering, Socio Fundador de Datais

➤ CONOCIMIENTOS RECOMENDADOS

El alumno aporta los conocimientos previos de R y Python, así como de análisis exploratorio de datos, técnicas de visualización, gestión de bases de datos, además de los conocimientos aportados en las asignaturas de matemáticas, estadística, econometría métodos de decisión del plan de estudios.

Fundamentos de Estadística y Matemáticas:

Antes de adentrarse en Data Mining, es fundamental que los estudiantes tengan una comprensión sólida de conceptos estadísticos básicos, como la media, la mediana, la desviación estándar y la distribución de probabilidad. También deben estar cómodos con matemáticas, especialmente álgebra y cálculo básico. Estos conocimientos les ayudarán a comprender los algoritmos y las métricas utilizadas en Data Mining.

Programación y Manejo de Datos:

Los estudiantes deben tener experiencia previa en programación, preferiblemente en un lenguaje como Python o R, ya que la implementación de algoritmos de Data Mining a menudo implica escribir código. Además, deben estar familiarizados con la manipulación de datos, incluyendo la carga, limpieza y transformación de conjuntos de datos.

Conceptos de Bases de Datos y SQL:

Tener una comprensión básica de bases de datos y SQL (Structured Query Language) es beneficioso, ya que gran parte de la minería de datos implica trabajar con bases de datos para extraer y analizar información. Los estudiantes deben saber cómo realizar consultas SQL simples para recuperar datos de manera eficiente.

➤ OBJETIVOS TEÓRICOS

Conocimientos Teóricos y Prácticos en Data Mining Supervisado:

Desarrollar una comprensión sólida de las técnicas y algoritmos de aprendizaje automático supervisado, incluyendo la clasificación y la regresión, para aplicarlos adecuadamente en el

análisis de datos etiquetados.

Capacidad para Predecir Resultados:

Adquirir la habilidad de utilizar datos etiquetados para predecir resultados desconocidos, aplicando modelos de aprendizaje automático supervisado. Esto incluye la capacidad para construir, entrenar, y evaluar modelos predictivos con precisión.

Capacidad para Clasificar Datos en Categorías Predefinidas:

Desarrollar la competencia para clasificar conjuntos de datos en categorías específicas, utilizando técnicas de aprendizaje supervisado. Esto implica entender la naturaleza de los datos, los requerimientos de etiquetado, y la implementación de algoritmos adecuados para la tarea.

Habilidad para Identificar y Aplicar Técnicas de Preprocesamiento e Ingeniería de Características:

Aprender a preparar los datos para el análisis mediante el preprocesamiento y la ingeniería de características, mejorando así la eficiencia y efectividad de los modelos de aprendizaje supervisado. Esto incluye la limpieza de datos, la selección y transformación de características, y la reducción de dimensionalidad cuando sea necesario.

➤ OBJETIVOS PRÁCTICOS

- Trabajo de Análisis Exploratorio de Datos (EDA) para Aprendizaje Supervisado:
 1. Estadísticas Descriptivas: Los estudiantes necesitan dominar las estadísticas descriptivas para analizar la distribución de los datos y las variables objetivo. Esto es crucial para entender cómo cada característica influye en la variable que se está intentando predecir o clasificar.
 2. Visualización de Datos: Es fundamental conocer técnicas avanzadas de visualización que permitan identificar patrones de relación entre las variables independientes y la variable dependiente, así como anomalías o outliers que podrían afectar el rendimiento del modelo.
- Trabajo en Selección y Transformación de Características:
 1. Conocimientos en Selección de Características: Entender cómo identificar y seleccionar las características más relevantes para los modelos supervisados es esencial. Esto incluye técnicas como el análisis de importancia de variables y métodos de selección de características basados en modelos.
 2. Ingeniería de Características: Desarrollar habilidades para crear nuevas características a partir

de los datos existentes que puedan mejorar la precisión de los modelos de aprendizaje supervisado.

- **Trabajo de Modelado Predictivo:**

1. **Técnicas de Clasificación y Regresión:** Los estudiantes deben aprender a aplicar y comparar distintos algoritmos de aprendizaje supervisado para clasificación y regresión, como Árboles de Decisión, SVM, Regresión Logística, y Redes Neuronales.
2. **Evaluación de Modelos:** Es crucial conocer métodos de evaluación específicos para el aprendizaje supervisado, como la curva ROC, el área bajo la curva (AUC), la precisión, el recall, y el F1 score, para determinar la eficacia de los modelos predictivos.

- **Trabajo Global de Aplicación de Modelos Supervisados:**

1. **Integración de Procesos de Data Mining Supervisado:** Los estudiantes deben ser capaces de realizar un proyecto completo que integre EDA, selección y transformación de características, y modelado predictivo, aplicando estos procesos para resolver un problema de negocio real.
2. **Selección de Modelos:** Aprender a elegir el algoritmo más adecuado para un problema específico, teniendo en cuenta las características de los datos y los objetivos del análisis.
3. **Comunicación de Resultados:** Desarrollar habilidades para presentar los resultados de manera efectiva, incluyendo la interpretación de los coeficientes del modelo, la importancia de las características, y las predicciones, con el fin de informar decisiones basadas en datos.

➤ **COMPETENCIAS QUE SE DESARROLLAN / RESULTADOS DE APRENDIZAJE**

Competencias Básicas y Generales

- **CB2** Que los estudiantes sepan aplicar sus conocimientos a su trabajo o vocación de una forma profesional y posean las competencias que suelen demostrarse por medio de la elaboración y defensa de argumentos y la resolución de problemas dentro de su área de estudio.
- **CB3** Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.
- **CB4** Que los estudiantes puedan transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado.
- **CG1** Resolver problemas de análisis de negocio en función del contexto y de los factores y variables más relevantes.
- **CG2** Analizar y sintetizar la información, hipótesis y variables más importantes de un libro, un tema, un artículo, un caso, etc.

Competencias Específicas

- **CE2** Comprender las herramientas estadísticas y econométricas para el análisis de variables económicas y empresariales.

- CE6 Saber manejar herramientas cuantitativas e informáticas para la toma de decisiones
- CE7 Conocer los fundamentos de la estadística multivariante y data mining, así como su aplicación al análisis de negocios.
- CE11 Capacidad para manipular y a modelizar patrones con el uso de la inteligencia espacial aplicada a negocio. (geolocalización).
- CE19 Adquirir un enfoque de gestión analítico y orientado al dato, ayudar a las compañías en su transformación data-driven. Así como el alumno adquirirá la capacidad para conocer y aplicar los principios fundamentales y técnicas básicas de los sistemas inteligentes y su aplicación práctica

Competencias Transversales

- CT1 Desarrollar el pensamiento crítico propio del espíritu universitario, así como la capacidad de analizar, argumentar e interpretar datos relevantes y complejos para poder integrarlos de manera sólida y solvente en la toma de decisiones.
- CT2 Utilizar con rigor y precisión el lenguaje oral y escrito, siendo capaz de transmitir información a un público tanto especializado como no especializado, teniendo en cuenta los diferentes contextos.

➤ CONTENIDO DEL PROGRAMA

1 Introducción al Machine Learning

Diferencias entre Modelos Supervisados y No Supervisados: Explicación de los conceptos fundamentales del aprendizaje automático, con énfasis en las características que distinguen a los modelos supervisados de los no supervisados.

2. Aplicaciones de los Modelos Supervisados: Casos de Éxito

Ejemplos Prácticos y Estudios de Caso: Presentación de aplicaciones exitosas de modelos supervisados en diversos sectores, como la salud, finanzas, marketing, y más, ilustrando cómo pueden resolver problemas complejos de predicción y clasificación.

3. Preprocesamiento e Ingeniería de Variables

Técnicas de Limpieza y Preparación de Datos: Enseñanza de métodos para la limpieza y el preprocesamiento de datos, así como estrategias avanzadas de ingeniería de variables para mejorar la precisión y eficacia de los modelos supervisados.

4. Técnicas de Selección de Características

Métodos para la Reducción de la Dimensionalidad: Introducción a técnicas específicas para la selección de características en el contexto supervisado, como la eliminación hacia atrás, selección hacia adelante y métodos basados en modelos.

5. Algoritmos de Clasificación

Desde Regresión Logística hasta Máquinas de Soporte Vectorial (SVM): Explicación y aplicación práctica de varios algoritmos de clasificación, incluyendo árboles de decisión, random forests, y redes neuronales, con ejemplos y ejercicios prácticos.

6. Algoritmos de Regresión

Regresión Lineal, Árboles de Decisión para Regresión, y más: Exploración de algoritmos para tareas de regresión, enseñando cómo modelar y predecir valores continuos a partir de datos.

7. Métodos Ensemble y Boosting

Boosting, Bagging, Random Forest, y Gradient Boosting: Detalle de cómo combinar múltiples modelos para mejorar la precisión y robustez de las predicciones, incluyendo técnicas como XGBoost y LightGBM.

8. Evaluación de Modelos y Ajuste de Hiperparámetros

Métricas de Rendimiento y Validación Cruzada: Descripción de las métricas de evaluación específicas para el aprendizaje supervisado y técnicas para el ajuste de hiperparámetros, como la búsqueda en cuadrícula y la búsqueda aleatoria.

9. Redes Neuronales y Deep Learning

Aplicaciones de Redes Neuronales en Aprendizaje Supervisado: Introducción a los fundamentos del deep learning y cómo se aplican las redes neuronales para tareas de clasificación y regresión, incluyendo ejemplos con TensorFlow y Keras.

10. Predicción de Series Temporales

Modelos para Datos Temporales: Exploración de técnicas y modelos específicos para la predicción de series temporales, abordando desafíos como la estacionalidad y tendencias.

11. Criterios de Evaluación de los Modelos Supervisados

Evaluación Integral del Rendimiento de Modelos: Discusión sobre cómo evaluar de manera efectiva los modelos supervisados, incluyendo consideraciones prácticas para garantizar que las evaluaciones sean justas y representativas.

12. Resumen y Perspectivas Futuras

Reflexión sobre el Aprendizaje Supervisado y su Futuro: Resumen de los conceptos clave y discusión sobre las tendencias emergentes en el aprendizaje automático supervisado, preparando a los estudiantes para los avances futuros en el campo.

➤ ACTIVIDADES FORMATIVAS

Actividades formativas

Denominación	Nº horas	% Presencialidad
Lecciones magistrales	30	100%
Resolución de problemas	45	50%
Exposiciones orales de trabajos	2	100%
Elaboración de informes y escritos	15	0%
Tutorías	2	100%
Trabajo autónomo	51,6	0%
Pruebas de conocimientos	4,4	100%

➤ CRITERIOS Y MÉTODOS DE EVALUACIÓN**Convocatoria Ordinaria 1a Matricula**

Los criterios que se aplican son los siguientes:

- Evaluación Continua: 50%
- Asistencia y participación activa: 10%
- Actividades individuales o grupales realizadas dentro o fuera del aula: 40%
- Examen Final: 50% (necesario tener mínimo un 4) para tener en consideración la evaluación continua)

Convocatoria Ordinaria a partir de 2ª Matricula y dispensa académica

El criterio de asistencia y participación se suprime. El alumno será evaluado con los siguientes parámetros en convocatoria ordinaria:

- Evaluación continua (50%): Actividades individuales realizadas dentro o fuera de clase planificadas e informadas a los alumnos con la suficiente antelación.
- Examen final (50%) (necesario tener mínimo un 4 para tener en consideración la evaluación continua)

Convocatoria Extraordinaria

- En caso de tener la evaluación continua aprobada (mayor o igual que 5), se aplicarán los mismos criterios que en la convocatoria ordinaria, manteniéndose la nota ya obtenida en la evaluación continua ordinaria.

- **Evaluación Continua: 50%**
- **Examen Final: 50% (necesario tener mínimo un 4 para tener en consideración la evaluación continua)**

- En caso de haber suspendido la evaluación continua a lo largo del curso, el alumno podrá recuperar la mitad del porcentaje de esta a través de pruebas o trabajos decididos por el profesor. La otra mitad del porcentaje de la evaluación continua no se puede recuperar y ponderará la nota obtenida en la evaluación continua ordinaria.

- **Evaluación Continua: 25%**
- **Evaluación Continua recuperable: 25% (a través de pruebas o trabajos)**
- **Examen Final: 50% (necesario tener mínimo un 4 para tener en consideración la evaluación continua)**

Calificación No presentado

La calificación de "No Presentado", en cada una de las dos convocatorias oficiales, no consume convocatoria.

Cuando el alumno se haya presentado a todas las pruebas de evaluación, o a una parte de las mismas, siempre que su peso en la nota final supere el 30%, aunque no se presente al examen final, no podrá ser calificado como No presentado en convocatoria ordinaria, siendo la nota final de la asignatura en convocatoria ordinaria la nota de la evaluación continua.

Se entenderá que un alumno se ha presentado a una prueba si asiste a la misma aunque la abandone una vez comenzada la misma.

La condición de "No Presentado" en la convocatoria extraordinaria estará ligada a la no asistencia al examen final.

TRATAMIENTO DE LA REDACCIÓN Y LAS FALTAS DE ORTOGRAFÍA

En la evaluación de los exámenes y trabajos escritos del estudiante se tendrá en cuenta su capacidad de redacción, manifestada en la exposición ordenada de las ideas, el correcto engarce sintáctico, la riqueza léxica y la matización expresiva. Se tendrá además en cuenta la propiedad del vocabulario, la corrección sintáctica, la corrección ortográfica (grafías y tildes), la puntuación apropiada y la adecuada presentación. En el caso de examen, los errores ortográficos sucesivos se penalizarán con un descuento de 0,25 cada uno, hasta un máximo de dos puntos.

➤ **BIBLIOGRAFÍA Y RECURSOS BÁSICOS**

"Introduction to Data Mining" por Pang-Ning Tan, Michael Steinbach, y Vipin Kumar. "Data Mining: Practical Machine Learning Tools and Techniques" por Ian H. Witten, Eibe Frank, y Mark A. Hall. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" por Trevor Hastie, Robert Tibshirani, y Jerome Friedman. "Mining of Massive Datasets" por Jure Leskovec, Anand Rajaraman, y Jeffrey D. Ullman. "Pattern Recognition and Machine Learning" por Christopher M. Bishop.

Recursos bibliográficos de la Universidad Villanueva: <https://biblioteca.villanueva.edu/>

➤ **HORARIO, TUTORÍAS Y CALENDARIO DE EXÁMENES**

1.-Los horarios podrán ser consultados en tiempo real en: Horarios

Aproximadamente una sesión por tema.

2.-El calendario de exámenes podrá ser consultado en el siguiente link sobre el general para Grados del curso 2024-25. Calendario Académico

Fechas y Tareas Clave: Iremos concretando en las próximas semana

Primera práctica EDA:

Segunda práctica Reducción Dimensional:

Tercera práctica Clustering:

Práctica Final:

Horarios de atención (tutorías): Jueves de 1830 a 1930

➤ **PUBLICACIÓN Y REVISIÓN DE LA GUÍA DOCENTE**

Esta guía docente se ha elaborado de acuerdo a la memoria verificada de la titulación.