

1. Datos Generales de la asignatura

Nombre de la asignatura:	Minería de Datos y Minería de Texto
Clave de la asignatura:	<u>Sólo para el caso de especialidades, el Instituto Tecnológico se apegará al Lineamiento para la Integración de Especialidades vigente. En los demás casos DGEST asignará las claves.</u>
SATCA¹:	2-4-6
Carrera:	Maestría en Sistemas Computacionales

2. Presentación**Caracterización de la asignatura**

La asignatura de Minería de Datos y Minería de Texto permite desarrollar soluciones de software para el apoyo a la toma de decisiones a partir de la explotación de datos históricos y el desvelamiento de hallazgos. La aplicabilidad de estas soluciones se asocia prácticamente a cualquier campo en el que se generen datos tanto semiestructurados como no estructurados, incluyendo texto en lenguaje natural.

La importancia de esta asignatura radica en la posibilidad de preparar al profesional de la Ingeniería en Sistemas Computacionales, la Ingeniería Informática y otras carreras afines para adentrarse en el mundo de la Ciencia de Datos y hacer frente a la demanda creciente del mercado en tal sentido.

En el primer tema se explica, de manera general, el modelo de proceso estandarizado de la Minería de Datos, incluyendo las fases de entendimiento del negocio, preprocesamiento, modelamiento, evaluación y despliegue, y se hace énfasis en importancia y en su adaptabilidad a los procesos de analítica de textos.

El segundo tema se enfoca en la fase de preprocesamiento del modelo de proceso estandarizado de la Minería de Datos y demuestra la aplicación de técnicas de limpieza, escalamiento y normalización de datos, portabilidad de tipos de datos, muestreo de datos y selección de características.

El tercer tema se enfoca en la fase de modelamiento de dicho modelo de proceso y demuestra la aplicación de técnicas de aprendizaje automático para clasificación,

¹ Sistema de Asignación y Transferencia de Créditos Académicos

regresión y agrupamiento en un lenguaje de programación en particular, para lo que se sugieren Python y R. Asimismo, permite introducirse en temáticas emergentes complementarias relacionadas con la explotación de datos no estructurados y de texto en lenguaje natural para la determinación de polaridades y sentimientos empleando técnicas de aprendizaje automático y de Procesamiento del Lenguaje Natural (PLN).

Finalmente, el cuarto tema demuestra la aplicación del proceso estandarizado de la Minería de Datos para la realización de análisis descriptivos y la construcción y evaluación de modelos de análisis predictivo a partir de textos y datos semiestructurados históricos como medio para la resolución de problemas del entorno. Para ello se sugieren Python y R.

La asignatura de Minería de Datos y Minería de Texto requiere como competencias previas habilidades y conocimientos básicos de Probabilidad y Estadística, programación estructurada y programación orientada a objetos, Álgebra lineal, Cálculo Diferencial y Matemáticas Discretas, entre otras.

Intención didáctica

El enfoque sugerido para la asignatura requiere que las actividades prácticas promuevan el desarrollo de habilidades para la resolución de problemas, tales como la selección, implementación y evaluación de las técnicas de aprendizaje automático para análisis predictivo más apropiadas dada la naturaleza de un problema a resolver; asimismo, este requiere que dichas actividades prácticas propicien procesos intelectuales como inducción-deducción y análisis-síntesis con la intención de generar una actividad intelectual compleja. Las actividades teóricas se consideran como actividades previas al tratamiento práctico de los temas. Por otro lado, en las actividades prácticas sugeridas, es conveniente que el profesor sólo guíe al estudiante en la construcción del conocimiento.

Se sugiere que en el último tema de la asignatura se desarrolle un proyecto final que permita integrar las competencias específicas desarrolladas a lo largo de todos los temas del temario a fin de realizar la selección, tratamiento y análisis descriptivo de datos empresariales históricos o conjuntos de datos públicos de interés científico y aplicabilidad en el entorno local y regional, la selección e implementación de técnicas de aprendizaje automático y redes neuronales artificiales aplicables al desarrollo de modelos de análisis predictivo y la extracción de conocimiento para su interpretación y evaluación. No obstante, idealmente, este proyecto se integraría con las distintas actividades prácticas realizadas durante todo el curso. Es recomendable que al inicio del curso el docente les hable a los estudiantes acerca del desarrollo de este proyecto final de asignatura, y que se establezcan las características y requisitos generales de este en dicho momento.

En el transcurso de la realización de las actividades programadas es muy importante que el estudiante entienda que está construyendo su quehacer futuro y que, en consecuencia, actúe de una manera profesional; de igual manera, se espera que el estudiante aprecie la importancia del conocimiento desarrollado y de los hábitos de trabajo, que desarrolle las capacidades de análisis y síntesis, la capacidad de investigación, la curiosidad, la puntualidad, el entusiasmo y el interés, la tenacidad, la flexibilidad y la autonomía.

Es necesario que el profesor preste atención y cuidado en los aspectos antes mencionados en el desarrollo de las actividades de aprendizaje de esta asignatura.

El diseño del temario de la asignatura, específicamente, el número de temas y subtemas de este, refleja la extensión y la profundidad con la que se propone que se aborde la asignatura durante un curso regular dentro del SNIT.

3. Competencia(s) a desarrollar

Competencia(s) específica(s) de la asignatura
Aprender los conceptos generales de la minería de datos, sus metodologías básicas y aplicaciones, así como también el descubrimiento de patrones y su estudio de tipos diferentes, en conjuntos de datos.

4. Competencias previas

Programa en algún lenguaje de programación.
Conoce álgebra.
Posee conocimientos básicos de probabilidad y estadística.
Conoce métodos de conteo (permutaciones y combinaciones).
Instala herramientas de software.

5. Temario

No.	Temas	Subtemas
1	Introducción a la minería de datos	1.1. Introducción a la Minería de Datos 1.2. Metodología de la Minería de datos 1.2.1. Entendimiento de los datos 1.2.2. Preparación de los datos 1.2.3. Modelamiento 1.2.4. Evaluación 1.2.5. Despliegue 1.3. Tendencias de la Minería de Datos
2	Preparación de los datos y arquitecturas de almacenamiento	2.1. Compilación de conjuntos de datos semiestructurados 2.2. Preprocesamiento y análisis descriptivo de datos semiestructurado 2.2.1. Portabilidad de tipos de datos 2.2.2. Limpieza de datos 2.2.2 Transformación y reducción de datos 2.3. Compilación de corpus de datos no estructurados. 2.4. Introducción al Procesamiento del Lenguaje Natural 2.4.1. POST tagging 2.4.2. Análisis de dependencias 2.4.3. Reconocimiento de entidades 2.4.4. N-gramas
3	Algoritmos y técnicas de minería de datos	3.1. Fundamentos teóricos del aprendizaje automático.

		3.2. Fundamentos teóricos de las redes neuronales artificiales. 3.2.1. Arquitecturas típicas 3.3. Regresión 3.3.2. Regresión lineal 3.3.3. Máquinas de vectores de soporte 3.4. Clasificación 3.4.2. K-vecinos más cercanos 3.4.3. Bosque aleatorio 3.5. Clustering 3.5.2. K-medias
4	Aplicación de la minería de datos	4.1. Análisis descriptivo 4.1. Modelado predictivo 4.2. Minería de opiniones 4.2.1. Clasificación de polaridad 4.2.2. Clasificación de emoción

6. Competencias de los temas

Introducción a la minería de datos	
Competencias	
Específica(s): Conocer e identificar las etapas y actividades del modelo de proceso estandarizado de la Minería de Datos. Conocer las tendencias actuales y futuras de la Minería de Datos. Genéricas: • Capacidad de abstracción, análisis y síntesis • Capacidad crítica y autocrítica • Habilidad para la búsqueda y análisis de información proveniente de fuentes diversas. • Capacidad de investigación • Capacidad creativa	• Realizar mapas mentales de las etapas y actividades del modelo de proceso estandarizado de la Minería de Datos. • Buscar propuestas de metodologías de Minería de Datos en literatura científica y en sitios web oficiales de empresas de tecnologías de la información y de la comunicación. • Comparar diversas metodologías estándares de Minería de Datos empleando cuadros comparativos. • Reconocer y discutir grupalmente las tendencias actuales y futuras de la Minería de Datos.
Preparación de los datos y arquitecturas de almacenamiento	
Competencias	
Específica(s): Identificar los componentes y las relaciones en diversas canalizaciones (<i>pipelines</i>) típicas de minería de datos. Conocer, seleccionar e implementar técnicas de preprocesamiento y de análisis descriptivo de datos. Genéricas: • Capacidad de abstracción, análisis y síntesis • Capacidad crítica y autocrítica • Habilidades en el uso de las tecnologías de la información y de la comunicación. • Habilidad para la búsqueda y análisis de información proveniente de fuentes diversas. • Capacidad de investigación • Capacidad para identificar, planear y resolver problemas.	• Elaborar diagramas que muestren los componentes y las relaciones de diferentes canalizaciones (<i>pipelines</i>) de minería de datos. • Elaborar cuadros comparativos de paquetes y funciones para preprocesamiento y para análisis descriptivo de datos en un lenguaje de programación típico. • Elaborar cuadros comparativos de paquetes y funciones para tareas de procesamiento del lenguaje natural en un lenguaje de programación típico. • Realizar estudios de casos de la identificación, selección e implementación de técnicas de preprocesamiento y de análisis descriptivo de datos. • Realizar estudios de casos de la

• Capacidad para tomar decisiones • Capacidad creativa • Capacidad de aplicar los conocimientos en la práctica.	identificación, selección e implementación de técnicas de procesamiento del lenguaje natural.
Algoritmos y técnicas de minería de datos	
Competencias	
Específica(s): Conocer los fundamentos teóricos del aprendizaje automático y de las redes neuronales artificiales, diferenciando las tareas o clases de problemas y los enfoques y arquitecturas principales. Conocer, seleccionar e implementar técnicas de aprendizaje automático y arquitecturas de redes neuronales para clasificación, regresión y agrupamiento. Genéricas: • Capacidad de abstracción, análisis y síntesis • Capacidad crítica y autocrítica • Habilidades en el uso de las tecnologías de la información y de la comunicación. • Habilidad para la búsqueda y análisis de información proveniente de fuentes diversas. • Capacidad de investigación • Capacidad para identificar, planear y resolver problemas. • Capacidad para tomar decisiones • Capacidad creativa • Capacidad de aplicar los conocimientos en la práctica.	• Elaborar infografías sobre las tareas o clases de problemas y enfoques de aprendizaje automático, así como sobre las principales arquitecturas de redes neuronales artificiales. • Elaborar cuadros comparativos de paquetes y funciones para aprendizaje automático de clasificación, regresión y agrupamiento en un lenguaje de programación típico. • Elaborar cuadros comparativos de paquetes y funciones para redes neuronales artificiales para clasificación, regresión y agrupamiento en un lenguaje de programación típico. • Realizar estudios de casos de la identificación, selección e implementación de técnicas de aprendizaje automático para clasificación, regresión y agrupamiento. • Realizar estudios de casos de la identificación, selección e implementación de arquitecturas de redes neuronales artificiales para clasificación, regresión y agrupamiento.
Aplicación de la minería de datos	
Competencias	
Específica(s): Reconocer áreas de oportunidad en el entorno local y regional para la aplicación de soluciones de minería de datos y de texto en la resolución de	• Desarrollar un proyecto que implique la selección, tratamiento y análisis de datos empresariales

problemas y la satisfacción de necesidades. Genéricas: • Capacidad de abstracción, análisis y síntesis • Capacidad crítica y autocrítica • Habilidades en el uso de las tecnologías de la información y de la comunicación. • Habilidad para la búsqueda y análisis de información proveniente de fuentes diversas. • Capacidad de investigación • Capacidad de trabajo en equipo • Capacidad para identificar, planear y resolver problemas. • Capacidad para tomar decisiones • Capacidad creativa • Capacidad de aplicar los conocimientos en la práctica.	semiestructurados o conjuntos de datos semiestructurados públicos y interés científico, la selección e implementación de técnicas de aprendizaje automático aplicables al desarrollo de modelos de análisis predictivo y la extracción de conocimiento para su interpretación y evaluación. • Llevar a cabo un proyecto que implique la compilación de un corpus de opiniones públicas de interés comercial o científico, su tratamiento mediante la selección e implementación de técnicas de Procesamiento del Lenguaje Natural y su análisis mediante la selección e implementación de arquitecturas de redes neuronales artificiales, que resulte en un modelo de clasificación de polaridades/sentimientos, que permita la extracción de conocimiento para su interpretación y evaluación.
--	--

7. Práctica(s)

Con el objetivo de vincular y fortalecer el aprendizaje del saber con el saber hacer se recomienda el desarrollo de las siguientes prácticas:

Recopilación y preprocesamiento de conjuntos de datos propios
 Predicción con conjuntos de datos históricos públicos
 Evaluación de modelos de predicción
 Clasificación de polaridad en comentarios procedentes de la web
 Clasificación de emoción en comentarios procedentes de la web

8. Fuentes de información

1. Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst* (1a ed.). Wiley.
2. Aggarwal, C. C. (2015). *Data Mining: The Textbook* (1st ed.). Springer Publishing Company, Incorporated.
3. Dean, J. (2014). *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. Wiley.
4. Fregly, C., & Barth, A. (2021). *Data Science on AWS: Implementing End-to-End, Continuous AI and Machine Learning Pipelines* (1a ed.). O'Reilly Media.
5. Larose, D. T. (2015). *Data Mining and Predictive Analytics* (2nd ed.). Wiley.
6. Moroney, L. (2017). *The Definitive Guide to Firebase: Build Android Apps on Google's Mobile Platform* (1a ed.). Apress.
7. Mukunthu, D., Shah, P., & Tok, W. H. (2019). *Practical Automated Machine Learning on Azure: Using Azure Machine Learning to Quickly Build AI Solutions* (1a ed.). O'Reilly Media.
8. Zaki, M. J., & Jr, W. M. (2020). *Data Mining and Machine Learning: Fundamental Concepts and Algorithms* (2a ed.). Cambridge University Press.
9. Rao D, & McMahan B. (2019). *Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning* (1a ed.). O'Reilly Media.
10. Liu B. (2020). *Sentiment Analysis: mining sentiments, opinions, and emotions* (2a. ed.). Cambridge University Press.