UNTREF | UNIVERSIDAD NACIONAL DE TRES DE FEBRERO

MAESTRÍA EN GENERACIÓN Y ANÁLISIS DE INFORMACIÓN ESTADÍSTICA

CICLO DE: **Seminarios electivos**

MATERIA: **Ciencia de datos: Tópicos de captura, análisis y visualización de información**

CARGA HORARIA: **32 horas**

AÑO: **2024** CUATRIMESTRE: **Segundo**

NOMBRE Y APELLIDO DE DOCENTE TITULAR**: Mg. Juan Manuel Fernández** NOMBRE Y APELLIDO DOCENTE ADJUNTO: **A definir**

CORRELATIVIDADES:

1-PRESENTACIÓN Y OBJETIVO DEL CURSO

Este curso tiene por finalidad dotar a quienes cursan la Maestría en Generación y Análisis de Información Estadística de conocimientos y competencias vinculadas a la disciplina de ciencia de datos, desde la captura de los datos, pasando por técnicas de análisis y generación de conocimiento hasta el proceso de presentación del conocimiento obtenido, ya sea a través de visualizaciones como de informes.

El objetivo principal del curso es que las y los maestrandos se apropien de los principales conceptos de ciencias de datos así como la aplicación de técnicas y herramientas de aprendizaje automático y visualización para el análisis de datos y consecuente toma de decisiones informadas.

OBJETIVOS PARTICULARES

Se procura que quienes realicen este curso se apropien de conocimiento y desarrollen competencias que les permitan:

- Comprender y desarrollar las distintas etapas del ciclo de vida de un proyecto de ciencia de datos.

- Adquirir conocimientos y competencias sobre herramientas y tecnologías vinculadas a la captura de datos desde grandes repositorios de información. - Conocer y aplicar algoritmos de aprendizaje automático para la explicación, predicción y clasificación de información.
- Desarrollar habilidades prácticas en la visualización efectiva de datos, utilizando herramientas y técnicas modernas para comunicar resultados de manera clara y comprensible.
- Aplicar el conocimiento adquirido en casos reales, mediante la resolución de problemas prácticos y la toma de decisiones informadas basadas en el análisis riguroso de datos.

2-DESARROLLO DE LOS CONTENIDOS

CONTENIDOS MÍNIMOS:

Introducción a la ciencia de datos. Modelos de procesos para la ciencia de datos. Captura y almacenamiento de información. Integración de datos. Bases de datos no convencionales. Aprendizaje automático supervisado: técnicas supervisadas y no supervisadas. Diseño e implementación de visualizaciones para grandes volúmenes de datos.

PROGRAMA ANALÍTICO

UNIDAD 1. Introducción a la ciencia de datos

Definición de Ciencia de Datos. Relación con la Estadística. Importancia y Aplicaciones. Modelos de Procesos en Ciencia de Datos. Fases de CRISP-DM. Pasos en el Proceso KDD. Tendencias y Futuro. Aplicaciones Prácticas.

UNIDAD 2. Captura y almacenamiento de datos

Obtención de datos históricos y datos en tiempo real. Utilización de APIs y herramientas para capturar datos. Extracción y representación de datos textuales. Extracción de datos de las redes sociales. Captura de información de la web. Web scraping y web crawling. Uso de expresiones regulares. Librerías y frameworks

específicos. Integración de fuentes de datos heterogéneas. Bases de datos no convencionales.

UNIDAD 3. Aprendizaje automático

Conceptos básicos de aprendizaje automático y modelado predictivo. Tipos de algoritmos: supervisados (regresión, clasificación) y no supervisados (agrupamiento, reglas de asociación). Evaluación de modelos: validación cruzada, matriz de confusión, curvas ROC, métricas de selección de modelos. Aplicaciones prácticas en análisis de datos y predicción de variables.

UNIDAD 4. Visualización de grandes volúmenes de datos

El proceso de la visualización y representación de datos. Formulación de preguntas. Tipos de datos. Datos complejos y grandes volúmenes de datos. Técnicas descriptivas multidimensionales. Herramientas avanzadas de visualización: gráficos interactivos, mapas de calor, gráficos de redes. Principios de diseño de visualización efectiva: claridad, simplicidad, integridad. Comunicación de resultados a audiencias no técnicas: storytelling con datos, dashboards interactivos.

3-BIBLIOGRAFÍA PRINCIPAL

UNIDAD 1. Introducción a la ciencia de datos

- Skiena, S. S. (2017). The data science design manual. Springer. - Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). Mining of massive data sets. Cambridge University Press.
- Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan Kaufmann.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

UNIDAD 2. Captura y almacenamiento de datos

- Ryan Mitchell. Web Scraping with Python: Collecting Data from the Modern

Web. O'Reilly Media, Inc., 2015.

- A. Silberschatz, H. Korth, S. Sudarshan, “Database System Concepts”, 7ma Edición, McGrawHill, 2019.
- Russell, M. A. (2019). Mining the social web: data mining Facebook, Twitter, LinkedIn, Google+, GitHub, and more. O'Reilly Media, Inc.
- Meier, A., & Kaufmann, M. (2019). SQL & NoSQL databases (pp. 123-142). Wiesbaden:: Springer Fachmedien Wiesbaden.
- Heydt, M. (2018). Python Web Scraping Cookbook: Over 90 proven recipes to get you scraping with Python, microservices, Docker, and AWS. Packt Publishing Ltd.

UNIDAD 3. Aprendizaje automático

- Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). Mining of massive data sets. Cambridge University Press.
- Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
- Han, J., Pei, J., & Tong, H. (2022). Data mining: concepts and techniques. Morgan Kaufmann.

UNIDAD 4. Visualización de grandes volúmenes de datos

- Wilke, C. O. (2019). Fundamentals of data visualization: a primer on making informative and compelling figures. O'Reilly Media.
- Subasi, A. (2020). Practical machine learning for data analysis using python. Academic Press.
- Knaflic, C. N. (2015). Storytelling with data: A data visualization guide for business professionals. John Wiley & Sons.
- Kirk, A. (2019). Data visualisation: A handbook for data driven design. Data Visualisation, 1-328.
- McCandless, D. (2014). Knowledge is beautiful. London: William Collins.

3b-BIBLIOGRAFÍA COMPLEMENTARIA

UNIDAD 1. Introducción a la ciencia de datos

- Daniel T. Larose. 2014. Segunda edición. Discovering Knowledge in Data: An Introduction to Data Mining.
- Aggarwal, C. C. (2015). Data mining: the textbook. Springer.

UNIDAD 2. Captura y almacenamiento de datos

- Sullivan, D. (2015). NoSQL for mere mortals. Addison-Wesley Professional.

UNIDAD 3. Aprendizaje automático

- Bramer, M., & Bramer, M. (2016). Data for data mining. Principles of data mining, 9-19.
- Daniel T. Larose. 2014. Segunda edición. Discovering Knowledge in Data: An Introduction to Data Mining.

UNIDAD 4. Visualización de grandes volúmenes de datos

- Whitney, H. (2012). Data insights: new ways to visualize and make sense of data. Newnes.
- Sievert, C. (2020). Interactive web-based data visualization with R, plotly, and shiny. Chapman and Hall/CRC.

4-METODOLOGÍA y MODALIDAD DE CURSADA

Las clases serán sincrónicas, con al menos un 50% de encuentros presenciales y en su totalidad de carácter teórico-práctico.

Inicialmente, se presentarán y discutirán los conceptos de los temas a abordar en ese encuentro, proponiendo con anterioridad la lectura de la bibliografía a quienes tomen el curso para propiciar su participación activa en la discusión. Por otra parte, en cada clase se plantean y resolverán casos prácticos de aplicación de los conceptos tanto en la

pizarra como en computadora.

Todas las presentaciones utilizadas en las clases se pondrán a disposición de los estudiantes en el aula virtual del curso.

La ejercitación en computadoras se realizará utilizando dos lenguajes de programación ampliamente empleados tanto en la industria como en la comunidad científica para la ciencia de datos como Python y R, complementados con diferentes frameworks y librerías aplicadas para las temáticas del curso.

Por otro lado, se empleará un gestor de bases de datos relacional como PostgreSQL y la interfaz pgAdmin IV y se presentarán algunas variantes de Bases de Datos NoSQL como (Neo4j y MongoDB, por ejemplo) para la integración de datos desde múltiples fuentes. En cuanto a los set de datos, se trabajará con conjuntos de datos reales provenientes de distintas fuentes.

5-REQUISITOS PARA LA CURSADA Y PROMOCIÓN

a- ASISTENCIA A CLASES

Se aplicará la normativa de la Universidad, con relación al presentismo y al cumplimiento de cualquier otro requisito que la misma imponga. Se requerirá cumplir con, al menos, un 75% de asistencia a todas las clases sincrónicas, independientemente que sean o no presenciales.

b- EVALUACIÓN

La evaluación del curso se basa en la realización de un trabajo final en el que deberán aplicar las competencias desarrolladas durante el curso, principalmente orientadas a la captura, análisis y visualización de datos.

Para alcanzar la **regularidad**, los estudiantes deberán presentar al finalizar el curso el problema a abordar, el diseño conceptual de la base de datos y el modelo relacional de la misma.

Para la **aprobación**, deberán presentar la implementación de la base y las consultas y/o procesos activos que permitan la generación de la información necesaria para dar respuesta al problema planteado.

5-ORGANIZACIÓN DE CLASES

#	Temario / Unidad	Modalidad		
		Presencial		Virtual
		Aula común	Aula laboratorio	
Clase 1	Unidad 1	4 hs		
Clase 2	Unidad 2			4 hs
Clase 3	Unidad 2		4 hs	

Clase 4 Unidad 3 4 hs

Clase 5	Unidad 3		4 hs	
Clase 6	Unidad 3			4 hs
Clase 7	Unidad 4			4 hs
Clase 8	Unidad 4		4 hs	

5-FECHA DE EXÁMEN O ENTREGA DE TRABAJO FINAL

Para alcanzar la **regularidad del curso**, los estudiantes deberán entregar antes de la finalización del curso el proyecto que abordarán como Trabajo Final. El proyecto deberá contener el problema a abordar, las fuentes de datos que utilizarán y los análisis a emplear o visualizaciones a realizar, al menos hipotéticamente.

UNTREF | UNIVERSIDAD NACIONAL DE TRES DE FEBRERO

Alcanzada la regularidad, para la **aprobación del curso** deberán presentar la implementación de la solución en alguna de las fechas de examen final. La presentación consistirá en la defensa del modelo, visualizaciones o solución implementada que permiten la generación de la información necesaria para dar respuesta al problema planteado.

a- PRIMERA FECHA: A confirmar

b- SEGUNDA FECHA: A confirmar

Encuentros virtuales los días lunes y miércoles de 19.00 a 22.00 horas Fecha de inicio: 5/08/24. Fecha de finalización: 28/08/24. Total de horas: 32

Fecha límite de pago de matrícula: 1 de agosto de 2024

ARANCEL: 96.000 pesos

Alumnos regulares, docentes y graduados de la Universidad Nacional de Tres de Febrero reciben una reducción arancelaria del 50%.

Informes e Inscripción
maestriaestadistica@untref.edu.ar