



Facultad de Ciencias
**Astronómicas
y Geofísicas**
UNIVERSIDAD NACIONAL DE LA PLATA

MODALIDAD POSGRADO

Carga Horaria Total: 150 horas (90 horas cátedra y 60 horas de trabajo autónomo)

Requisitos establecidos para su aprobación:

- Asistencia del 80% como mínimo a clases teórico/prácticas
- Aprobación en término y en forma de todas las evaluaciones vinculadas a los ejercicios de los Trabajos Prácticos. Realización de tres exposiciones.
- Examen integrador.
- Desarrollo de una actividad suplementaria vinculada con la temática del curso y con el plan de tesis del doctorando.

PROGRAMA DE LA ASIGNATURA FUNDAMENTOS DE CIENCIAS DE DATOS

CARRERA: Licenciatura en Astronomía, Geofísica y Licenciatura en Meteorología y Ciencias de la Atmósfera.

CARGA HORARIA: 64hs (4hs semanales).

CARÁCTER: Cuatrimestral, 1er cuatrimestre (y/o 2do cuatrimestre según disponibilidad).

RESPONSABLE/S: Prof. Dra. Erika Gularte y Prof. Dr. Gustavo Baume.

CONTACTO: erika@fcaglp.unlp.edu.ar

MODALIDAD: Grado

CORRELATIVAS: Computación ó Análisis Numérico y Programación.

MESA EXAMINADORA: Prof. Dra. Erika Gularte /Prof. Dr. Gustavo Baume /Prof. Dr. Carlos Feinstein.

OBJETIVOS

La ciencia de datos es un campo interdisciplinario que utiliza técnicas estadísticas, computacionales y matemáticas para extraer información y conocimientos a partir de los datos. El objetivo general del curso consiste en adquirir los conceptos, las técnicas y las herramientas fundamentales de la ciencia de datos.

Por su parte, los objetivos específicos delineados para este curso son los siguientes:

- Desarrollar habilidades para recopilar, limpiar y manipular datos para su análisis.
- Aplicar distintas técnicas de modelado, mediante aprendizaje automático, a fin de identificar patrones y tendencias en los datos. Así como también, para predecir resultados y tomar decisiones informadas.
- Adquirir habilidades de visualización, tanto sea para analizar los datos como para comunicar los resultados de manera clara y efectiva.

Además, un objetivo primordial, que surge en el transcurso de todo el curso, es realizar aplicaciones prácticas de la ciencia de datos en áreas de interés para la Astronomía, la Geofísica, la Meteorología y las Ciencias de la Atmósfera.

DESCRIPCIÓN GENERAL

La información reduce nuestra incertidumbre sobre algún aspecto de la realidad. Desde la antigüedad, la información se ha recopilado y organizado para asistir en la toma de decisiones, permitiendo de esta forma que dichas decisiones se encuentren más adecuadas para la resolución un determinado problema. A finales del siglo XX, con la irrupción de Internet, la información, en formato digital, se diseminó en todos los ámbitos. En las últimas décadas, el almacenamiento, organización y recuperación de la información se ha automatizado gracias a los sistemas de bases de datos. Así, en la actualidad se dispone de un gran volumen y variedad de información que se encuentra en numerosas bases de datos digitales y en otras fuentes. En este último caso se puede incluir el instrumental científico, los modelos teóricos y/o simulaciones computacionales. Cabe señalar que la información



involucrada puede pertenecer a diferentes dominios y usualmente se encuentra en continuo crecimiento.

Distintas herramientas se han utilizado para el análisis de los datos. En particular, las herramientas de "procesamiento analítico en línea" (OLAP = On-Line Analytical Processing) soportan cierto análisis descriptivo y de síntesis que permiten transformar los datos originales en otros datos agregados o cruzados de manera sofisticada. Pero estas herramientas no generan reglas, patrones, pautas, es decir, conocimiento que pueda ser aplicado a otros datos. Por otra parte, la estadística ha sido la primera ciencia que considera los datos como su materia prima. Ella permite inferir patrones a partir de los datos, ya sea utilizando una modelización paramétrica o no paramétrica. No obstante, algunas metodologías estadísticas generalmente no funcionan de forma apropiada cuando se los aplica al volumen de datos actuales (cientos de tablas, millones de registros, varios gigabytes y/o una alta dimensionalidad). Además, se presentan dificultades con el manejo de algunos tipos especiales de datos, tales como, atributos nominales con muchos valores, datos textuales, multimedia, etc. Por otro lado, ciertos paquetes estadísticos no se integran bien con los sistemas de información actuales. Todos estos problemas y limitaciones de las aproximaciones tradicionales han hecho surgir la necesidad de una nueva generación de herramientas y técnicas aptas para soportar la extracción de conocimiento útil desde la información disponible.

De esta forma, la disciplina de la "ciencia de datos" no aparece por el desarrollo de tecnologías esencialmente diferentes a las preexistentes, sino que se crea, en realidad, por la aparición de las nuevas necesidades, en particular, las nuevas características de los datos, tanto sea en volumen como en su tipología. Así como también, el requerimiento de establecer procesos automatizados y herramientas de manejo sencillo para el tratamiento de grandes volúmenes de datos. Adicionalmente, otro de los desafíos de la ciencia de datos consiste en establecer técnicas de visualización adecuadas, para que tanto la información analizada como la descubierta sea más comprensible para el usuario. Por consiguiente, la ciencia de datos tiene como finalidad fundamental la extracción de conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos heterogéneos almacenados en distintos formatos. Más específicamente, se busca la extracción de patrones, describir tendencias y regularidades, y predecir comportamientos de los datos. Dicho proceso, automático o semiautomático, consta de varias fases: la recopilación de datos, mediante el uso de almacenes de datos o de manera directa; la preparación de datos, haciendo uso de la visualización, agregación, limpieza o transformación; la minería de datos, empleando técnicas descriptivas o predictivas; la evaluación y mejora de modelos, a través de la validación cruzada, combinación o análisis de costes, y finalmente; la difusión y uso del conocimiento extraído.

El presente curso se encuentra orientado a que los estudiantes desarrollen destrezas para utilizar apropiadamente las diversas técnicas existentes en cada una de las fases de la extracción de conocimiento a partir de diferentes tipos de datos. Dentro de los retos buscados, los estudiantes deben ser capaces de reconocer las metodologías y técnicas apropiadas para resolver distintos problemas, conociendo tanto sus alcances como sus limitaciones. El curso se ha distribuido en ocho capítulos, contemplando los tópicos de ciencia de datos de mayor interés para los estudiantes de nuestra Facultad. En el Capítulo I se presentan los conceptos y herramientas básicas. El Capítulo II aborda los distintos tipos



de datos y su adquisición/recopilación a través de diversas fuentes. Tanto el análisis inicial como el análisis exploratorio de los datos se presentan en el Capítulo III. En la etapa inicial se trata de comprender los datos antes de realizar un análisis más detallado, identificando errores y problemas que puedan afectar la calidad de los resultados del análisis posterior. En el análisis exploratorio de los datos se investigan y resumen sus características principales. Esto permite tomar decisiones informadas sobre cómo proceder con el análisis y modelado subsiguiente. El Capítulo III se complementa con la etapa del proceso en la que se combinan y unifican diferentes fuentes de datos para crear un conjunto de datos coherente y completo. Además, se analiza la selección de los atributos a ser considerados para el análisis. Luego, en los Capítulos IV y V, se aborda el aprendizaje automático, enfocado en el desarrollo de modelos que permitan a las computadoras aprender a partir de los datos y realizar tareas específicas sin ser programadas explícitamente. Se presenta el aprendizaje no supervisado, en el que un modelo busca patrones y relaciones en un conjunto de datos sin la ayuda de etiquetas o respuestas previas, y el aprendizaje supervisado, en el que un modelo es entrenado utilizando un conjunto de datos previamente etiquetados o clasificados para predecir la variable objetivo para nuevas entradas. En los dos capítulos siguientes, se abordan dos tópicos específicos de interés. Por un lado, en el Capítulo VI se detalla el aprendizaje profundo a través de las redes neuronales. Siendo éste una rama del aprendizaje automático que se caracteriza por procesar datos complejos con una precisión y velocidad mucho mayor que los métodos tradicionales. Por el otro, en el Capítulo VII se analiza el caso particular de las series temporales en el contexto de la ciencia de datos. Se presentan diversos modelos específicos. Finalmente, debido a la dinámica que representan los contenidos de la asignatura, el Capítulo VIII se deja sin especificar. El mismo abarca diferentes tópicos particulares y/o relevantes que se irán actualizando según la oferta o demanda. Cabe señalar que, cada uno de los temas a desarrollar, cuenta con un soporte adecuado de visualización y con una práctica orientada a los temas de interés de un astrónomo/a, un geofísico/a y/o un meteorólogo/a.

CONTENIDO TEMÁTICO

- I. INTRODUCCIÓN. Evolución histórica. Conceptos generales. El proceso KDD y la minería de datos. Aprendizaje automático. Lenguajes de programación. Software.
- II. EL DATO. Tipos de datos. Visualización de los datos. Formatos de archivos. Bases de datos. Números aleatorios.
- III. PREPROCESAMIENTO. Análisis inicial de los datos. Análisis exploratorio de los datos. Limpieza e integración de los datos. Selección y transformación de atributos.
- IV. APRENDIZAJE NO SUPERVISADO. Concepto básico. Métricas. Métodos de agrupamiento. Reglas de asociación. Comparación de métodos. Evaluación.
- V. APRENDIZAJE SUPERVISADO. Concepto básico. Partición de datos. Métodos de clasificación. Métodos de regresión. Comparación de métodos. Evaluación.
- VI. APRENDIZAJE PROFUNDO. Concepto básico. Redes Neuronales. Tipos de redes neuronales. Funcionamiento de la red. Visualización de la red. Aplicaciones.
- VII. SERIES DE TIEMPO. Concepto básico. Tipos de datos temporales. Estructura interna de la serie de tiempo. Modelos lineales y no lineales. Selección y validación de modelos. Modelos con memoria de largo y corto plazo.
- VIII. MISCELÁNEAS.



BIBLIOGRAFÍA

- Géron A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2022). O'Reilly Media.
- Gray A., J.T. VanderPlas, A.J. Connolly, Ž. Ivezić. *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data* (2020). Princeton University Press.
- Han J., M. Kamber, J. Pei. *Data Mining: Concepts and Techniques 3rd Edition* (2011). Elsevier, Morgan Kaufmann.
- Hastie T., R. Tibshirani, J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2013). Springer.
- Müller A.C., S. Guido. *Introduction to Machine Learning with Python: A Guide for Data Scientists* (2016). O'Reilly Media.
- Orallo J.H., M.J. Ramírez Quintana, C. Ferri Ramírez. *Introducción a la Minería de Datos* (2004). Pearson Prentice Hall.
- Press William H., S.A. Teukolsky, W.T. Vetterling, B.P. Flannery. *Numerical Recipes: The Art of Scientific Computing 3rd Edition* (2007). Cambridge University Press.
- Witten I.H., E. Frank, M.A. Hall, C.J. Pal. *Data Mining Practical Machine Learning Tools and Techniques* (2007). Morgan Kaufmann.

La bibliografía presentada se complementará con información disponible en diferentes páginas web de sitios especializados. Adicionalmente, la naturaleza de la asignatura hace necesaria una actualización y mejora regular de la bibliografía involucrada que será brindada oportunamente a los alumnos.

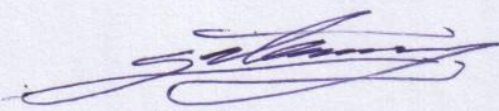
MODALIDAD DE APROBACIÓN

La asignatura puede aprobarse por promoción, para lo cual se establecen los siguientes requisitos: la asistencia al 80% de las clases teórico-prácticas, la aprobación de un parcial teórico-práctico debiendo alcanzarse un mínimo de siete (7) puntos y un coloquio final donde se procura verificar que el alumno haya aprehendido e integrado los conceptos fundamentales de la materia.

Aquellos alumnos que no optaran por el régimen de promoción o no lograran cumplimentar los requisitos de la misma, pueden acogerse a la modalidad de cursada regular usual con la aprobación del examen parcial teórico-práctico con un mínimo de cuatro (4) puntos y un examen final.



Prof. Dra Erika Gularte



Prof. Dr. Gustavo Baume