**UNIVERSIDAD NACIONAL DE SANTIAGO DEL ESTERO**

**FACULTAD DE CIENCIAS FORESTALES**

**CURSO DE POSGRADO:**

# Minería de datos en ciencias ambientales

PROFESOR RESPONSABLE:

Priscilla G. Minotti

2024

SANTIAGO DEL ESTERO

**Nombre del Curso:** Minería de datos en ciencias ambientales

**Profesor Responsable:** Priscilla G. Minotti

**Profesores Colaboradores:**

**Carga Horaria:** 40 horas

**Créditos**: 4

**Fecha/horario:** 4, 5, 19, 25 y 26 de octubre, de 18 a 21h los viernes, y sábados de 9 a 12h

**Horario:** teórico- practicas (prácticas de gestión de datos con R) por la mañana de 8:30 a 12:30, el primer día de clase por la tarde se definirá horarios con alumnos para las clases de la tarde.

**Modalidad:** híbrida sincronico

## 1. Marco conceptual e importancia actual/Fundamento

Estamos en la Era de la Información, con una generación continua de datos ambientales. Sin embargo, se utiliza menos del 0,5% de lo que se produce, y muchas veces dichos datos se usan una única vez. La minería de datos o *data mining*, constituye un componente clave de la generación de conocimiento a partir de datos. Es el proceso de identificación automatizada de información relevante extraída de grandes volúmenes de datos, con el objetivo de descubrir patrones y tendencias, estructurando la información obtenida de un modo comprensible para su posterior utilización. Las técnicas de minería de datos se apoyan en la estadística, la gestión de bases de datos y el desarrollo y aplicación de algoritmos de aprendizaje automático (*machine learning*), temas que no suelen ser abordados de manera conjunta en las disciplinas ambientales. El curso busca dar herramientas para un mejor aprovechamiento de la información ambiental con foco en disciplinas ambientales como Ingeniería ambiental y Ecología.

## 2. Objetivos

El objetivo general del curso es introducir enfoques de minería de datos de utilidad en Ciencias Ambientales.

Como objetivos específicos se espera que los participantes aprendan conceptos de ciencia de datos, se familiaricen con técnicas de minería de datos en entorno del lenguaje R, y apliquen alguno de los enfoques vistos en clase a sus áreas de estudio o trabajo.

**3. Contenidos**

Temas Teóricos (cada tema corresponde a una clase)

1. Introducción. Minería de datos y Big Data. Proceso de minería de datos. Patrones y modelos. Generalizacion. Tareas de minería de datos. Minería de datos y aprendizaje automático (machine learning). Mineria de datos y ética. Ejemplos de aplicación. Explicación de las actividades de evaluación.

2. Datos. Tipos de datos. Calidad de datos. Preprocesamiento: Agregación, Muestreo, reducción de dimensionalidad, Selección de variables (atributos o features). Variables derivadas, discretización, binarizacion, Transformaciones. Medidas de similitud, disimilitud. Análisis exploratorio.

3. Conceptos básicos de Clasificación. Esquema general. Clase objetivo y variables predictoras. Algoritmo básico de Arboles de decision.. Hiperparametros. Poda para evitar sobreajuste.Comparación de modelos. Ejemplos de aplicaciones.

4. Conceptos básicos de Agrupamiento. Tipos de agrupamientos. Tipos de grupos. Numero de grupos. K-means (K-Medias). Agrupamiento jerárquico aglomerativo. DBScan. Evaluación de grupos. Ejemplos de aplicaciones.

5. Clasificación con Redes Neuronales Artificiales. Perceptron simple: arquitectura, algoritmo básico. Parámetros. Perceptron multicapa. Características generales de las RNA. Conceptos de Aprendizaje profundo (Deep Learning). Ejemplos de aplicaciones.

6. Ensambles de modelos de clasificación. Concepto. Métodos de construcción de ensambles de clasificadores. Descomposición sesgo-varianza. Baging. Boosting. Random Forest. Comparaciones entre ensambles de modelos. Métodos de balanceo de datos. Medidas de performance agregadas. Ejemplos de aplicaciones.

7. Agrupamiento con redes neuronales artificiales: SOM (Self Organized Maps). Ordenamiento topologico. Arquitectura. Inicializacion. Selección de objetos. Asignacion. Actualizacion de peos. Terminación. Visualizacion de salidas. Limitaciones. Ejemplos de aplicaciones.

8. Otros enfoques de Agrupamiento y Clasificación. Clasificación: vecino más cercano, SVM (suport vector machine), basados en reglas, Bayes ingenuo., regresión logística. Agrupamiento: difuso, basado en modelos mixtos, en densidad. Otras áreas de minería de datos. Minería de textos. Minería de grafos. Análisis de canasta y patrones frecuentes. Minería de datos con Big Data.

**Trabajos prácticos** (cada TP corresponde a un día distinto)

TP1. Problemas ambientales, tipos de datos y tareas de minería de datos aplicables. Ejemplos de trabajos publicados y en blogs. Video Tutorial 1: Introducción a RStudio y RStudio Cloud. Introducción al lenguaje R. Asignación de trabajos grupales a presentar en clase y explicación del TP .

TP2. Preparación de un dataset ambiental. Video Tutorial 2: Instalación e introducción al tidyverse. El paquete readr para leer datos tabulares. El paquete ggplot2 y adicionales para hacer gráficos estadísticos. Los paquetes dplyr y tidyr para limpiar y reorganizar datos tabulares.

TP3. Clasificación mediante un árbol de decisión. Video tutorial 3. Modelo de árboles de decisión con el paquete rpart.plot.

TP4. Agrupamiento mediante k-means. Video tutorial 4-1. Modelo de agrupamiento con el paquete ClusterR. Video tutorial 4.2. Cluster jerárquico con la función hclust del paquete stats.

TP 5. Clasificación con Redes Neuronales Artificiales. Pregunta de interés ambiental. Variable objetivo y variables predictoras relevantes. Selección de muestras d entrenamiento y testeo. Video Tutorial 5-1. Redes neuronales someras con los paquetes nnet y neuralnet. Video Tutorial 5-2. El paquete unet.

TP6. Clasificación con Random Forest. Video Tutorial 6. El paquete randomForest.

TP7. Agrupamiento con SOM. Video Tutorial 7. El paquete Kohonen.

TP 8. Presentaciones grupales sobre papers de aplicaciones ambientales. Video tutorial sobre los paquetes caret y tidymodels.

TP 9. Revision de flujos de trabajo de minería de datos para áreas de interés de los participantes.

TP10. Presentacion de propuestas de trabajo final.

**4. Evaluación**

Cada clase se acompaña de ejercitación sobre el tema teórico visto y se complementa con un video tutorial, para seguir fuera de clase.

El curso tiene dos instancias de evaluación.

Para certificar la asistencia al curso deberán dar una presentación grupal de un trabajo publicado sobre la aplicación de algún método de minería de datos a un problema ambiental y

aquellos que quiera aprobar el curso deberá presentar una propuesta de análisis para el trabajo final.

Para aprobar el curso de posgrado, deberán entregar y presentar el desarrollo de un trabajo final sobre datos propios, que según su complejidad puede ser realizado de manera grupal o individual.

**5. Infraestructura necesaria**

Para temas teóricos: espacio-aula de audio video comunicación

Para prácticos: opciones 1) Sala de computación con: R y RStudio en últimas versiones instalados, acceso a Internet y permisos para descargar e instalar paquetes nuevos; directorio con permisos de lectura y escritura para guardar datos y proyectos de R; posibilidad de tener pantalla de proyección para clase de TP sincrónica por Zoom o equivalente. 2) Cada alumno deberá proveer su propio equipo de computación para realizar las practicas, con los mismos requerimientos que en sala de computación.

**6. Bibliografía básica**

(Los libros en pdf o epub serán facilitados por la docente).

Tan PN, Steinbach M, Karpatne A, y Kumar V. 2019. Introduction to Data Mining. Global Edition. 2nd. Edition. Pearson Education Limited. eBook ISBN 13: 978-0-273-77532-4.

Witten IA, Frank E, Hall E, Pal C. 2016. Data Mining: Practical Machine Learning Tools and Techniques. 4thEdition. Morgan Kaufmann Series in Data Management Systems. Materiales complementarios disponibles en https://www.cs.waikato.ac.nz/ml/weka/book.html.

RStudio. Primers. https://rstudio.cloud/learn/primers

Jamsa K. 2021. Introduction to Data Mining and Analytics with machine learning in R and python. Jones & Bartlett Learning. ISBN: 9781284218688. Datasets de practica y primeros dos capitulos en https://www.jblearning.com/catalog/productdetails/9781284180909#productInfo

**7. Estrategias de enseñanza (obligatorio p/ educación a distancia o con estrategias de hibridación)**

El curso es de modalidad híbrida sincronica.

La Universidad Nacional de Santiago del Estero tiene implementado un Sistema de Educación a Distancia (SIED), regulado por Res. Ministerial E2641 (2017) y creado mediante CS 178 (2018).

En Contenidos Teóricos se presenta el plan de aprendizaje. Los Trabajos Prácticos serán de tipo virtual y se señalan los distintos contenidos que serán tutoreados por el docente, videotutoriales para que cada alumno tenga información de carácter práctico adicional, y también se especifican la presentación de trabajos grupales y finales individuales.