Pontificia Universidad Católica Argentina

FACULTAD DE INGENIERÍA Y CIENCIAS AGRARIAS

| | |
|---|---|
| **CARRERA: Licenciatura en Ciencias de Datos** | **CURSO LECTIVO: 2024** |
| **CÁTEDRA: Exploración de Datos** | **CURSO: 1º año - 2º semestre** |
| **DURACIÓN: Semestral** | **Hs. TOTALES: 48 Hs. Reloj Totales** |
| **SEMANAS: 16** | **Hs. TEÓRICAS: 16 Hs. Reloj Totales**<br>**Hs. PRÁCTICAS: 32 Hs. Reloj Totales** |

**PROFESOR PROTITULAR:** Quintero, Antonio

## 1. OBJETIVOS DE LA ASIGNATURA

Que los alumnos logren:

- Familiarizarse con las ciencias de datos.
- Reconocer la importancia de la disciplina en diversos ámbitos: el ámbito social, el empresarial, el profesional, el académico, el gubernamental, etc.
- Incorporar el alcance de las ciencias de datos.
- Comprender la perspectiva histórica, el desarrollo de las ciencias y técnicas asociadas a las ciencias de datos, darle sentido al presente de la práctica y vislumbrar las tendencias a futuro.
- Comprender cómo en ciencias de datos se aplica el método científico.
- Diferenciar correlación de relación causal. Familiarizar el sesgo estadístico de manera de poder llevar este concepto a la cotidianidad: falacias y fake news
- Conocer lo esencial de los principales modelos de ciencias de datos

## 2. UNIDADES TEMÁTICAS

### 1. Introducción.

Datos: Definición de “Dato”. Tipos de Datos. Datos estructurados. Datos semi-estructurados. Datos no estructurados. Alcance de las Ciencias de Datos. Ciencias Asociadas. Aplicaciones en la Industria y en otras Ciencias. Perspectiva histórica.

Ciencia de datos vs. machine learning vs. inteligencia artificial. Características de los datos: extracción de patrones significativos, construcción de modelos representativos, combinación de disciplinas, algoritmos de aprendizaje. Campos asociados indispensables: estadística descriptiva, visualización exploratoria, aplicaciones de procesamiento analítico en línea (OLAP), prueba de hipótesis, ingeniería de datos, ingeniería empresarial.

Resolución de problemas de programación: Algoritmos. Tipo abstracto de datos, Abstracción de datos. Análisis de un problema. Diseño y pseudocódigo de un programa: elección de una estructura de datos, modularización, interacción entre

módulos, precondición y postcondición. Implementación (codificación) de un programa: parámetros por valor, parámetros por referencia, parámetros por variable, variables globales. Verificación de un programa. Descripción formal de tipos de datos. Lenguaje de lógica de primer orden: axiomas, sintaxis, semántica, símbolos, término, fórmula, invariantes.

## 2. Estadística Descriptiva.

Estadística y datos. Observación, población y muestra. Variables: definición, cuantitativas, cualitativas, discretas, continuas. Escala: nominal, ordinal, métrica. Agrupamiento de datos: variables agrupadas, variables categóricas, variables binarias. Recopilación de Datos: encuesta, datos experimentales, datos de observación, datos primarios y secundarios. Creación de un conjunto de datos: observaciones vs. variables, transformaciones. Medidas de frecuencia: datos discretos, frecuencia absoluta, frecuencia relativa, datos métricos agrupados, función de distribución acumulativa empírica (ECDF) para variables ordinales y variables métricas. Representación gráfica de los datos: gráficos de barras, gráficos circulares, histogramas, gráficos de densidad de kernel. Medidas de Tendencia Central: media aritmética, media aritmética ponderada, mediana, cuantiles, QQ-plots, moda, media geométrica, media armónica. Medidas de dispersión: rango, rango intercuartílico, desviación absoluta, error cuadrático medio, varianza, desviación estándar, varianza para datos agrupados, teorema de la descomposición de la varianza, coeficiente de variación. Diagramas de caja. Medidas de concentración. Curva de Lorenz. Coeficiente Gini. Asociación de dos variables: Tabla de contingencia para datos discretos, distribuciones de frecuencia (conjuntas, marginales y condicionales), representación gráfica de dos variables nominales u ordinales, independencia, frecuencias esperadas. Medidas de asociación de dos variables: variables discretas, coeficiente $x^2$ de Pearson, coeficiente V de Cramer, coeficiente de contingencia C. Representación gráfica de dos variables: emparejamiento de variables, gráficos de dispersión, coeficiente de correlación, rangos de Spearman. Medidas para pares concordantes y discordantes.

## 3. Regresión.

Regresión lineal múltiple. Modelo lineal. Método de mínimos cuadrados: estimador, modelo de regresión, línea de regresión ajustada, signo de beta, i-ésimo valor ajustado, residual. Bondad de ajuste. Regresión lineal con una covariable binaria. Regresión Logística.

## 4. Introducción al modelado.

Modelos de aprendizaje supervisados, no-supervisados, semi-supervisados. Uso de los modelos. Modelo predictivo y descriptivo. Agrupamiento (Clustering) y agrupamiento predictivo. Conjunto de datos de entrenamiento y prueba. Modelos geométricos: separabilidad lineal, transformaciones lineales, distancia euclidiana, distancia Manhattan, clasificador vecino-más-cercano(k-NN), clasificador k-means, máquinas de vectores de soporte (SVM). Modelos probabilísticos: probabilidad apriori, probabilidad aposterirori, valores perdidos, función de verosimilitud, Bayes, regla de decisión máxima a posteriori (MAP), regla de decisión de máxima verosimilitud (ML), odds posteriori, modelo generativo, verosimilitud marginal. Modelos lógicos: árbol de características, superposición de reglas, algoritmos de aprendizaje de árboles. Modelos de agrupación y de calificación. Aprendizaje supervisado de modelos predictivos. Ruido: de instancia, de etiqueta. Características. Clasificación Binaria. Evaluación del desempeño de la clasificación: tabla de contingencia, matriz de confusión, métricas. Clase mayoritaria. Sensibilidad y especificidad vs. Precisión y recall. Puntuación y ranking. Margen y función de perdida, criterio de información de Akaike (AIC).

## 5. Características.

Tipos de características: cálculos sobre características, características categóricas, ordinales y cuantitativas, estructura de las características. Transformación de características: umbrales, discretización, normalización, calibración, características incompletas. Construcción de características y selección: transformación de matrices y descomposición. Algoritmos de selección de características para clasificación y

regresión: Pruebas de chi-cuadrado, algoritmo de mínima redundancia y máxima relevancia (MRMR), F-Test, puntuaciones laplacianas, algoritmos ReliefF y RReliefF, selección secuencial de características utilizando criterios personalizados. Codificación de variables categóricas en variables numéricas: codificación activa, codificación media (codificación objetivo), ponderación de la evidencia, codificación de exclusión, codificación ordinal, codificación hash. Balance: sobremuestreo y submuestreo de caracteristicas.

**6. Modelos de Conjunto.**
Bagging y Random Forest. Boosting. Mapeo: bias, varinza y margenes. Meta-aprendizaje.

## 3. BIBLIOGRAFÍA

### 3.1 BIBLIOGRAFÍA GENERAL OBLIGATORIA


- Introduction to Statistics and Data Analysis with Exercises, Solutions and Applications in R Springer 2st edition. Christian Heumann, Michael Schomaker, Shalabh. 2023. ISBN-10: 3031118324
- Machine learning. Cambridge University Press. 1st edition. Peter Flash. 2012. ISBN-10: 1107422221.
- Introduction to Machine Learning. The MIT Press. 4th edition. Ethem Alpaydin. 2020. ISBN- 10: 0262043793


## 4. METODOLOGÍA

El curso está organizado en 8 unidades temáticas divididas en encuentros de 4 hs cátedras semanales, a realizarse en formato presencial. La modalidad adoptada para el dictado será **teórico- práctica.** En las clases se presentarán los temas de cada unidad, proponiendo espacios de intercambio con el docente y entre los estudiantes a partir de consignas específicas. Se facilitará material de lectura obligatoria y complementaria para complementar la comprensión de las unidades.

## 5. EVALUACIONES Y CRITERIOS PARA LA APROBACIÓN

La aprobación de la materia estará supeditada al cumplimiento de la condición de asistencia exigida por la Universidad, la aprobación de todas las actividades prácticas y la aprobación del examen integrador.

Los trabajos prácticos podrán ser individuales o grupales, debiéndose cargar a través de la plataforma de Entornos Virtuales de Aprendizaje en tiempo y forma, otorgándose una única instancia de revisión y recuperación. Las actividades prácticas deberán contar con su aprobación para acceder a la instancia de evaluación final.

Para los trabajos prácticos y la evaluación final se realizarán sesiones de consultas individuales y grupales, haciendo además puesta en común general si el caso lo requiera. A los estudiantes que presenten dificultades se les observará y se los guiará para resolver el conflicto.

La instancia de recuperación está prevista para estudiantes que no hayan aprobado el examen integrador o que hayan estado ausentes.

Criterios de Evaluación:
- Respeto de las consignas presentadas.
- Resolución correcta de los problemas planteados.

- Adecuada respuesta a los contenidos teóricos.

## 6. CRITERIOS y MODALIDAD PARA LA EVALUACIÓN DEL EXAMEN FINAL

El examen final consiste en una evaluación oral y escrita, presencial e individual, donde el alumno deberá demostrar conocimientos teóricos y prácticos. El examen final se diferencia en que abarca todos los temas del programa. Los ejercicios prácticos tendrán un carácter integrador, articulando los distintos contenidos vistos en la materia. Finalmente, en las preguntas teóricas se pretende que el alumno demuestre un conocimiento profundo de los temas, relacionando conceptos entre sí.

Criterios de Evaluación:

- Respeto de las consignas presentadas.
- Adecuada respuesta a los contenidos teóricos.
- Relación de conceptos pertinente.
- Resolución correcta de los problemas planteados.
- Fundamentación bibliográfica de los temas.