Universidad de los Andes

Departamento de Física

# Introducción a la Ciencia de Datos

Nombre del curso: Introducción a la Ciencia de Datos
Prerrequisito: FISI 2028 (Métodos Computacionales)
Créditos: 3 créditos pregrado. 4 créditos posgrado.
Código del curso: FISI XXX
Unidad académica: Departamento de Física
Periodo académico: 202010
Horario (magistral): XXX
Horario (laboratorio - Sección 1): XXX

---

Nombre profesor magistral: Jaime Ernesto Forero Romero
Correo electrónico: je.forero@uniandes.edu.co
Horario y lugar de atención:

---

## I Introducción

La ciencia de datos (Data Science) se encuentra hoy en día en todas las áreas técnicas y científicas, dentro y afuera del ámbito académico. Esto se debe en gran parte a que la capacidad de procesar grandes cantidades de datos en computadoras de alto rendimiento ha disminuido en costo monetario y en complejidad.

El curso de *Introducción a la Ciencia de Datos* presenta estas posibilidades computacionales a estudiantes de diferentes disciplinas científicas. Para esto se propone profundizar sus conocimientos en dos áreas: métodos de descripción estadística de datos y la implementación de algoritmos para extraer patrones presentes en diferentes tipos de datos.

Se asume que los estudiantes de este curso ya tienen conocimientos básicos en métodos computacionales equivalentes al nivel del curso Métodos Computacionales (FISI-2028). El lenguaje de programación será Python.

## II Objetivos

El objetivo principal del curso es presentar algoritmos y técnicas básicas para:

- visualizar datos,
- describir datos con modelos estadísticos,
- clasificar conjuntos de datos con algoritmos.

## III Competencias a desarrollar

Al finalizar el curso, se espera que el estudiante esté en capacidad de:

- describir y clasificar datos con modelos lineales
- aplicar algoritmos de reducción de dimensionalidad de datos
- clasificar datos con redes neuronales.

## IV Contenido por semanas

**Semana 1.**

Presentación del curso. Probabilidad. Densidades de probabilidad. Valores esperados y covarianzas. Probabilidades Bayesianas. Ajuste de curvas Bayesiano.

**Semana 2.**
Teoría de la decisión de la información. Medidas de clasificación. Inferencia y decisión. Funciones de pérdidad para regresión. Entropía relative e información mutua.

**Semana 3**
Distribuciones de probabilidad. Variables binarias. Variables multinomiales. La distribución gaussiana.

**Semana 4**
Métodos de muestreo. Métodos Básicos. Markov Chain Monte Carlo. Gibbs sampling. Hamiltonian Montecarlo.

**Semana 5**
Visualizando datos. Análisis exploratorio de datos con Pandas.

**Semana 6**
Algebra lineal. Visualización de operaciones matriciales. Factorización de matrices. Autovalores y Autovectores. Descomposición en autovalores.

**Semana 7**
Modelos lineales para regresión. Descomposición Bias-Varianza. Regresión lineal bayesiana. Comparación de modelos Bayesiana.

**Semana 8**
Modelos lineals para clasificación. Funciones discriminantes. Modelos probabilísticos generativos. Modelos discriminantes probabilísticos.

**Semana 9**
Árboles de decisión y bosques aleatorios.

**Semana 10**
K-means clustering.

**Semana 11**
Gaussian Mixture Models.

**Semana 12**
Support Vector Machines.

**Semana 13**
Principal Component Analysis.

**Semana 14**
Redes Neuronales. Feed-forward Networks. Entrenamiento de redes. Backpropagation.

**Semana 15**
Redes Neuronales. Matriz Hessiana. Regularización en redes neuronales.

**Semana 16**
Redes Neuronales Convolucionales.

## V Metodología

En las sesiones magistrales, luego de presentar un resumen de los conceptos teóricos, se hará énfasis en la práctica computacional. Para que esto funcione es necesario que los estudiantes estudien el tema correspondiente **antes de cada clase** siguiendo las lecturas preparatorias recomendadas.

## VI Criterios de evaluación

Las componentes que reciben calificación en la Magistral (en paréntesis su contribución a la nota definitiva) son las siguientes:

- Asistencia (20 %). Cada asistencia a clase cuenta como una nota de 5.0 y una falta como 0.0. El promedio de esas notas será la nota de asistencia. Si hay **seis** o más fallas no justificadas durante todo el semestre esta nota es cero (0.0).

- Ejercicios (20 % cada uno). En cada clase hay un ejercicio para entregar. Cada ejercicio tiene dos partes. La primera se publica al menos un día antes de la clase y debe resolverse por fuera de la magistral. La segunda se publica y se resuelve durante la magistral. Durante el semestre el profesor eligirá a su discreción cuatro (4) de estos ejercicios para ser calificados.

## VII Bibliografía

Bibliografía principal:


- *Pattern Recognition and Machine Learning*. C. M. Bishop, Springer, 2006.
- *The Data Science Manual*. S. S. Skienna, Springer, 2017.
- *Python Data Science Handbook*. J. VanderPlas, O'Reilly, 2016.
- *An Introduction to Statistical Learning with Applications in R*, G. James, D. Witten, T. Hastie, R. Tibshirani, Springer, 2015