**SP-3068: Fundamentos para el aprendizaje de máquinas (aprendizaje estadístico)**

**Nombre del Programa:** Programa de Posgrado en Ciencias Cognoscitivas

**Plan de Estudios al que pertenece el curso:** Maestría Académica

**Modalidad:** Teórico-práctico

**Tipo de entorno:** Presencial

**Horas semanales:** 3

**Profesor que lo imparte:** Marcelo Araya Salas PhD [(marcelo.araya@ucr.ac.cr](mailto:(marcelo.araya@ucr.ac.crñ); sitio web: <https://marce10.github.io/>)

**Sitio web del curso:** <https://marce10.github.io/aprendizaje_estadistico_2024>

**Horas de consulta:** Martes 4 pm (oficina 2, Centro de Investigación en Neurociencias)

**Justificación**

El aprendizaje estadístico, también conocido como aprendizaje de máquinas o automático o "machine learning," es una rama de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender patrones y hacer predicciones para tomar decisiones basadas en datos. En lugar de seguir instrucciones explícitas para realizar una tarea, estos algoritmos identifican patrones en los datos y usan estos patrones para mejorar su desempeño en tareas específicas. El aprendizaje estadístico se utiliza en una amplia variedad de aplicaciones, desde el reconocimiento de voz hasta el análisis de grandes conjuntos de datos y la automatización de procesos industriales. En este curso los estudiantes podrán conocer los fundamentos y las técnicas básicas del aprendizaje estadístico para responder preguntas de investigación en Ciencias Cognoscitivas. Al iniciar el curso, se espera que las y los estudiantes conozcan los aspectos básicos de la estadística descriptiva e inferencial y el manejo básico del lenguaje de programación y análisis estadístico R.

**Objetivo General**

Capacitar a los estudiantes en los fundamentos, historia y diversas aplicaciones del aprendizaje de máquinas en el contexto de las Ciencias Cognoscitivas.

**Objetivos Específicos**

Al finalizar el curso, los estudiantes deberán ser capaces de:

* Describir y explicar los conceptos y métodos principales del aprendizaje estadístico aplicados a las Ciencias Cognoscitivas.
* Diferenciar entre los tipos de aprendizaje estadístico: supervisado, no supervisado, semi-supervisado y por reforzamiento.
* Comprender los fundamentos teóricos y los supuestos de técnicas como la regresión, las redes neuronales, los árboles de decisión y el análisis de conglomerados.
* Implementar, en el lenguaje de programación R, las técnicas desarrolladas a lo largo del curso.
* Seleccionar la técnica más adecuada en función del problema práctico o pregunta de investigación en las diversas áreas de las Ciencias Cognoscitivas.

### Contenidos

#### Introducción al Aprendizaje Estadístico

* Introducción a la estadística y su relación con el aprendizaje estadístico
* Diferencias y similitudes entre inteligencia artificial, aprendizaje estadístico y ciencia de datos
* Aplicaciones del aprendizaje estadístico en las Ciencias Cognoscitivas

#### Introducción a R

* Instalación y configuración de R y RStudio
* Estructuras de datos en R: vectores, matrices, listas y data frames
* Manipulación de datos con funciones base R
* Introducción a la visualización de datos con gráficos base R
* Escribir funciones y usar loops en R

#### Simulación de Datos con Patrones Predefinidos

* Generación de datos simulados en R
* Creación de patrones específicos para análisis
* Implementación en R y prácticas con simulación de datos

#### Regresión Lineal

* Modelo de regresión lineal
* Supuestos del modelo de regresión lineal
* Evaluación del modelo: R², error estándar, test F
* Implementación en R y prácticas con modelos lineales simples

#### Regresión Lineal Múltiple

* Modelo de regresión lineal múltiple
* Supuestos adicionales y diagnóstico del modelo
* Regresión con interacciones y variables categóricas
* Implementación en R y prácticas con modelos múltiples

#### Métodos de Clasificación

* Regresión logística
* Análisis discriminante lineal y cuadrático
* K-vecinos más cercanos (KNN)
* Implementación en R y prácticas con métodos de clasificación

#### Métodos de Remuestreo

* Validación cruzada
* Bootstrap
* Implementación en R y prácticas con técnicas de remuestreo

#### Selección y Regularización de Modelos

* Subconjunto, Ridge y Lasso
* Comparación de modelos y criterios de selección
* Implementación en R y prácticas con técnicas de regularización

#### Modelos de Árboles de Decisión

* Introducción a los árboles de decisión
* Árboles de clasificación y regresión
* Implementación en R y prácticas con árboles de decisión

#### Modelos de Ensamblado

* Bagging y Random Forests
* Boosting
* Implementación en R y prácticas con modelos de ensamblado

#### Máquinas de Soporte Vectorial

* Introducción a las SVMs
* Kernels y clasificación no lineal
* Implementación en R y prácticas con SVMs

#### Redes Neuronales y Deep Learning

* Introducción a las redes neuronales
* Perceptrón y neurona sigmoide
* Arquitectura de redes neuronales
* Algoritmo de propagación hacia atrás
* Implementación en R y prácticas con redes neuronales

#### Redes Neuronales y Deep Learning Avanzado

* Modelos de deep learning
* Redes convolucionales y recurrentes
* Implementación en R y prácticas avanzadas

#### Análisis de Conglomerados

* Introducción al análisis de conglomerados
* K-means y jerárquico
* Implementación en R y prácticas con análisis de conglomerados

#### Reducción de Dimensionalidad

* Análisis de componentes principales (PCA)
* Análisis de discriminante lineal (LDA)
* Implementación en R y prácticas con técnicas de reducción de dimensionalidad

**Método**

Se utilizará la plataforma de Mediación Virtual de la Universidad de Costa Rica como un recurso complementario para el acceso a los *scripts*, la entrega de las tareas y la comunicación entre el profesor y el estudiantado. Las sesiones teóricas y prácticas del curso serán presenciales. El metodo utilizado combina clases magistrales, demostraciones con software de algunos de los procedimientos de análisis de datos, y ejercicios prácticos que los y las estudiantes resolverán. Cuando se realicen los ejercicios prácticos, los resultados se discutirán en la clase para propiciar la interacción entre docente y estudiantes.

**Evaluación**

**Tareas (60%)**

Cada estudiante, de manera individual, deberá realizar 3 tareas con un valor de 20% cada una. En estas tareas, el estudiantado pondrá en práctica los conocimientos y destrezas abordados durante las sesiones presenciales del curso. Las tareas están diseñadas para evaluar la comprensión y aplicación de los conceptos teóricos y prácticos vistos en clase. Las tareas pueden incluir, pero no se limitan a:

- Implementación de algoritmos de aprendizaje estadístico en R.

- Análisis de datos utilizando técnicas de visualización y preprocesamiento.

- Resolución de problemas específicos utilizando modelos predictivos y de clasificación.

- Comparación y evaluación de diferentes métodos de aprendizaje estadístico.

**Trabajo Final (40%)**

El trabajo final tiene un valor de 40% y consiste en un proyecto integrador donde los estudiantes aplicarán todos los conocimientos adquiridos durante el curso para resolver un problema práctico. A continuación se detalla la estructura y los requisitos del trabajo final:

***1. Selección del Problema (10%)***

*- Los estudiantes deben elegir un problema de investigación o un caso práctico relevante en el contexto de las Ciencias Cognoscitivas.*

*- El problema seleccionado debe permitir la aplicación de técnicas de aprendizaje estadístico vistas en el curso.*

***2. Recopilación y Preprocesamiento de Datos (10%)***

*- Recopilación de un conjunto de datos adecuado para abordar el problema seleccionado.*

*- Realización de un preprocesamiento completo de los datos, incluyendo limpieza, transformación y visualización inicial.*

***3. Desarrollo e Implementación de Modelos (40%)***

*- Selección de al menos tres técnicas de aprendizaje estadístico diferentes para abordar el problema.*

*- Implementación de los modelos seleccionados en R, con una explicación detallada de los supuestos y parámetros utilizados.*

*- Evaluación comparativa de los modelos implementados utilizando métricas adecuadas (e.g., precisión, recall, F1-score, RMSE).*

***4. Análisis y Discusión de Resultados (20%)***

*- Interpretación de los resultados obtenidos de cada modelo.*

*- Discusión sobre las ventajas y desventajas de cada técnica aplicada en el contexto del problema seleccionado.*

*- Identificación de posibles mejoras y futuras direcciones de investigación.*

***5. Presentación y Reporte Final (20%)***

*- Elaboración de un reporte escrito que incluya introducción, metodología, resultados, discusión y conclusiones.*

*- Preparación de una presentación oral de 20-30 minutos (incluyendo preguntas) donde se expongan los principales hallazgos y se responda a preguntas del instructor y compañeros.*

El trabajo final permitirá a los estudiantes demostrar su capacidad para integrar y aplicar de manera crítica y creativa los conceptos y técnicas aprendidas a lo largo del curso. Además, les proporcionará una experiencia práctica que refuerce sus habilidades en el uso de R y en la resolución de problemas reales mediante el aprendizaje estadístico.

### Contenidos por sesión (3 horas semanales)

#### Día 1: Introducción a R para el Aprendizaje Estadístico - Parte 1

* Introducción al aprendizaje estadístico
* Instalación de R y RStudio
* Navegación en el entorno de RStudio
* Estructuras de datos básicas: vectores, matrices y listas
* Operaciones básicas y funciones en R
* Práctica con R: Crear y manipular vectores y data frames. Uso de funciones básicas
* ***Lectura:*** *Badillo et al (2020). An introduction to machine learning.* <https://ascpt.onlinelibrary.wiley.com/doi/full/10.1002/cpt.1796>

#### Día 2: Introducción a R para el Aprendizaje Estadístico - Parte 2

* Manejo de data frames y matrices
* Filtrado, ordenamiento y resumen de datos
* Visualización básica de datos con gráficos base
* Creación de gráficos personalizados
* Práctica con R: Manipulación de data frames, gráficos simples y personalizados

#### Día 3: Simulación de Datos con Patrones Predefinidos

* Introducción a la simulación de datos y su importancia
* Generación de datos con distribuciones específicas (normal, uniforme, etc.)
* Creación de datos con correlaciones y estructuras de dependencia
* Generación de datos categóricos y con ruido controlado
* Práctica con R: Simulación de conjuntos de datos para diferentes casos de estudio
* ***Lectura:*** *Roediger et al (2001). Factors that determine false recall: A multiple regression analysis.*
* <https://link.springer.com/content/pdf/10.3758/bf03196177.pdf>

#### Día 4: Regresión Lineal Simple

* Conceptos básicos de regresión lineal simple
* Ajuste de modelos y interpretación de coeficientes
* Evaluación del modelo: R² y error cuadrático medio (MSE)
* Diagnóstico de supuestos y multicolinealidad
* Tamaño de muestra y poder estadístico
* Práctica con R: Construcción y evaluación de modelos de regresión lineal simple

#### Día 5: Regresión Múltiple e Interacciones

* Introducción a la regresión múltiple
* Modelado con variables categóricas mediante variables ficticias
* Interacciones entre variables y su interpretación
* Evaluación y diagnóstico de modelos avanzados
* Práctica con R: Análisis de regresión múltiple e interacciones

#### Día 6: Regresión Logística

* Introducción a la clasificación y problemas de clasificación
* Regresión logística, uso e interpretación
* Predicciones a partir de un modelo
* Regresión multinomial
* Análisis de función discriminante
* Práctica con R: Implementación de modelos de clasificación y evaluación de desempeño
* ***Lectura:*** *Chen et al (2023). Identifying the top determinants of psychological resilience among community older adults during COVID-19 in Taiwan: A random forest approach.* <https://www.sciencedirect.com/science/article/pii/S2666827023000476>

#### Día 7: Métodos de Remuestreo y Evaluación de Modelos (Parte I)

* Validación cruzada y técnicas de remuestreo
* Uso de validación cruzada para evaluar modelos
* Medidas de evaluación: matriz de confusión, precisión, recall, índice F1
* Práctica con R: Implementación de técnicas de validación cruzada

#### Día 8: Evaluación de Modelos (Parte II)

* Validación cruzada y división de conjuntos de datos
* Curvas ROC y AUC: interpretación y uso
* Análisis de errores y ajuste de modelos
* Comparación y selección de modelos
* Práctica con R: Evaluación y comparación de modelos con diferentes métricas

#### Día 9: Métodos Basados en Árboles

* Árboles de decisión: construcción e interpretación
* Random forest: fundamentos y aplicaciones
* Evaluación de modelos de árboles y comparación con otros métodos
* Práctica con R: Construcción y evaluación de modelos basados en árboles
* Choi et al. 2020. ***Introduction to machine learning, neural networks, and deep learning.*** <https://tvst.arvojournals.org/article.aspx?articleid=2762344>

#### Día 10: Explorar Espacios Multidimensionales: Reducción de Dimensionalidad

* Introducción a la reducción de dimensionalidad y su necesidad
* Análisis de Componentes Principales (PCA): teoría y aplicación
* Visualización de datos en espacios reducidos
* Comparación con otros métodos como t-SNE
* Práctica con R: Aplicación de PCA y visualización de resultados

#### Día 11: Redes Neuronales y Deep Learning - Introducción

* Estructura de una red neuronal: neuronas, capas y activación
* Entrenamiento de redes neuronales: forward y backpropagation
* Introducción a Deep Learning y redes neuronales profundas
* Aplicaciones y casos de uso en el mundo real
* Práctica con R: Construcción de una red neuronal simple

#### Día 12: Redes Neuronales y Deep Learning - Aplicaciones Avanzadas

* Capas convolucionales y redes neuronales convolucionales (CNNs)
* Redes neuronales recurrentes (RNNs) y LSTM
* Técnicas de optimización y regularización
* Casos de estudio y aplicaciones en visión por computadora y procesamiento de lenguaje natural
* Práctica con R: Implementación de una CNN básica y una RNN
* Yarkoni & Westfall. 2017. *Choosing prediction over explanation in psychology: Lessons from machine learning.* <https://journals.sagepub.com/doi/pdf/10.1177/1745691617693393>

#### Día 13: Regularización y Generalización

* Conceptos de sobreajuste y subajuste
* Regularización: Lasso, Ridge y Elastic Net
* Prácticas para evitar el sobreajuste en modelos complejos
* Práctica con R: Aplicación de técnicas de regularización y ensamble

#### Día 14: Aprendizaje No Supervisado: Clustering

* Conceptos básicos de clustering y su importancia
* Método k-means: algoritmos y aplicaciones
* Clustering jerárquico: construcción de dendrogramas
* Evaluación de clusters: índice de Silhouette y coeficiente de Rand
* Práctica con R: Realización de análisis de clustering en datos de ejemplo

#### Días 15 y 16: Presentaciones del Proyecto Final

* Presentación del proyecto incluyendo una descripción detallada de los análisis de datos

### Bibliografía de Referencia

Alpaydin, E. (2020). **Introduction to Machine Learning.** Un recurso introductorio que cubre los conceptos clave y algoritmos en el aprendizaje estadístico, accesible para estudiantes con poca o ninguna experiencia previa. Enlace: [https://www.mitpress.mit.edu/books/introduction-machine-learning](https://www.mitpress.mit.edu/books/introduction-machine-learning" \t "/home/m/Documents\\x/_blank)

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam‐Thong, T., Siebourg-Polster J., Steiert B., & Zhang, J. D. (2020). **An introduction to machine learning.** Introduccion general a los metodos de aprendizaje estadístico mas robustos y otros comúnmente utilizados. Enlace: <https://ascpt.onlinelibrary.wiley.com/doi/full/10.1002/cpt.1796>

Bishop, C. M. (2006). **Pattern Recognition and Machine Learning.** Un libro esencial que abarca desde los fundamentos hasta las técnicas avanzadas en reconocimiento de patrones y aprendizaje estadístico. Enlace: [https://www.springer.com/gp/book/9780387310732](https://www.springer.com/gp/book/9780387310732" \t "/home/m/Documents\\x/_blank)

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). **Introduction to machine learning, neural networks, and deep learning.** Enlace: <https://tvst.arvojournals.org/article.aspx?articleid=2762344>

Goodfellow, I., Bengio, Y., & Courville, A. (2016). **Deep Learning.** Este libro es una referencia esencial para entender los conceptos fundamentales y avanzados de las redes neuronales y el deep learning. Enlace: [https://www.deeplearningbook.org](https://www.deeplearningbook.org" \t "/home/m/Documents\\x/_blank)

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2021). **An Introduction to Statistical Learning.** Proporciona una introducción accesible a los métodos estadísticos y de aprendizaje estadístico con aplicaciones prácticas en R. Enlace: [https://www.statlearning.com](https://www.statlearning.com" \t "/home/m/Documents\\x/_blank)

Hastie, T., Tibshirani, R., & Friedman, J. (2009). **The Elements of Statistical Learning: Data Mining, Inference, and Prediction.** Un recurso exhaustivo que cubre una amplia gama de métodos de aprendizaje estadístico y computacional, incluyendo teoría y aplicaciones. Enlace: [https://hastie.su.domains/ElemStatLearn/](https://hastie.su.domains/ElemStatLearn/" \t "/home/m/Documents\\x/_blank)

Hannes Rosenbusch, Felix Soldner, Anthony M. Evans, Marcel Zeelenberg. (2024). **Supervised machine learning methods in psychology: A practical introduction with annotated R code**. Introduccion general al aprendizaje estadistico supervisado en R. Enlace: <https://compass.onlinelibrary.wiley.com/doi/full/10.1111/spc3.12579>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). **An Introduction to Statistical Learning with Applications in R.** Un texto fundamental que cubre una amplia gama de técnicas de aprendizaje estadístico y computacional, con ejemplos prácticos en R. Enlace: [https://www.statlearning.com](https://www.statlearning.com" \t "/home/m/Documents\\x/_blank)

Kuhn, M., & Johnson, K. (2013). **Applied Predictive Modeling**. Este libro proporciona una guía práctica para la construcción de modelos predictivos utilizando R, con un enfoque en la preparación de datos y la selección de modelos. Enlace: [https://www.springer.com/gp/book/9781461468486](https://www.springer.com/gp/book/9781461468486" \t "/home/m/Documents\\x/_blank)

Murphy, K. P. (2012). **Machine Learning: A Probabilistic Perspective.** Ofrece una visión integral del aprendizaje estadístico desde una perspectiva probabilística, con numerosos ejemplos y ejercicios prácticos. Enlace: [https://mitpress.mit.edu/books/machine-learning](https://mitpress.mit.edu/books/machine-learning" \t "/home/m/Documents\\x/_blank)

Shalev-Shwartz, S., & Ben-David, S. (2014). **Understanding Machine Learning: From Theory to Algorithms.** Una introducción teórica sólida a los principios y algoritmos del aprendizaje estadístico. Enlace: [https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/](https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/" \t "/home/m/Documents\\x/_blank)

Wickham, H., & Grolemund, G. (2017). **R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.** Aunque centrado en tidyverse, es útil para comprender el flujo de trabajo en R y cómo aplicar principios de limpieza y visualización de datos. Enlace: [https://r4ds.had.co.nz](https://r4ds.had.co.nz" \t "/home/m/Documents\\x/_blank)

Yarkoni, T., & Westfall, J. (2017). **Choosing prediction over explanation in psychology: Lessons from machine learning.** Este articulo trata de las ventajas que puede traer el uso de herramientas de aprendizaje estadístico en la investigación en psicología. Enlace: <https://journals.sagepub.com/doi/pdf/10.1177/1745691617693393>