XS-4110: INTRODUCCION AL ANALISIS MULTIVARIADO

PROGRAMA

I SEMESTRE 2022

	GRUPO 01	**GRUPO 02**
Docente:	Ricardo Alvarado Barrantes.	Shirley Rojas Salazar.
Correo:	estad.ucr@gmail.com	ucrsrs@gmail.com
Teléfono:	84021263	87737058
Clases:	K-J: 1:00-2:50 pm	K-J: 1:00-2:50 pm
Consulta:	L: 4:00-6:00 pm	M: 9:00-11:00 am
Zoom:	https://udecr.zoom.us/j/83091352745	https://udecr.zoom.us/j/89709609762
Materiales:	https://www.dropbox.com/sh/wkbqfqe7edygpwg/AAB4z0XOdCb40MDDoS9yEi-ya?dl=0	

1. Descripción

Curso introductorio de técnicas estadísticas multivariadas para estudiantes de cuarto año del Bachillerato de Estadística, impartido con un enfoque teórico-práctico.
Además de los conocimientos teóricos se brindará al estudiante la posibilidad de aplicar los métodos mediante el uso de lenguaje estadístico R en trabajos de investigación.

- Requisitos: **XS-3310 Teoría Estadística y XS-2130 Modelos de Regresión Aplicados**
- Correquisitos: XS-4410 Práctica Profesional I
- Horas: 4 horas semanales
- Créditos: 4

2. Objetivo General

Aplicar las técnicas básicas, gráficas y cuantitativas, del análisis multivariado para la resolución de problemas que involucren varias variables y múltiples casos.

3. Objetivos Específicos

Al finalizar el curso el estudiante tendrá criterio y conocimiento básico para:

- Identificar los métodos multivariados y las técnicas de clasificación apropiados para su aplicación en situaciones específicas.
- Interpretar los resultados de los análisis realizados con cada método para la obtención de conclusiones y la toma de decisiones.
- Determinar el número de componentes principales para la reducción de la dimensionalidad de un grupo de variables.

- Identificar el tipo de distancia a calcular para el agrupamiento adecuado según el nivel de medición de las variables disponibles.
- Evaluar el desempeño de un modelo de clasificación para la selección del método más apropiado.
- Aplicar los métodos de ensamblaje de modelos para mejorar la eficiencia en la clasificación.
- Validar un modelo de clasificación para evitar sobreajuste.

4. Contenidos

I. *Análisis de componentes principales (PCA)*
1.1 Características y usos de los componentes principales. 1.2 Construcción: a) Valores y vectores propios. b) Cálculo de los puntajes en los componentes principales. c) Uso de covarianzas o correlaciones. 1.3 Representación gráfica: biplot. 1.4 Número de componentes principales. a) Variancia explicada. b) Criterios para decidir el número de componentes. 1.5 Evaluación de resultados: a) Reproducción de matriz de variancias. b) Correlación entre componentes y variables originales.
II. *Análisis de agrupamientos y escalamiento multidimensional (MDS)*
2.1 Distancias: a) Entre individuos (variables continuas, nominales, mezclas). b) Entre grupos (vecino más cercano, vecino más lejano, salto promedio). 2.2 Selección de variables para el análisis / Estandarización. 2.3 Agrupamientos jerárquicos: a) Algoritmo. b) Representación (dendograma). 2.4 Métodos de k-medias y k-medoides: a) Algoritmo. b) Selección del número de clústers. 2.5 Validación: número de clústers. 2.6 Presentación de resultados: mapas de calor. 2.7 Escalamiento multidimensional.

III. Técnicas de *clasificación*
3.1 Modelos de clasificación: a) Regresión logística binomial y multinomial b) Análisis discriminante c) K-vecinos más cercanos d) Máquinas de vectores de soporte (SVM) 3.2 Métodos basados en árboles: a) Arboles de decisión b) Agregación de Bootstrap (bagging) c) Bosques aleatorios (random forest) d) Boosting e) BART 3.3 Evaluación del modelo: a) Métricas de desempeño. b) Validación cruzada.

5. Metodología

- El curso seguirá la modalidad bimodal (virtual- presencial) usando la plataforma Zoom para las lecciones, las cuales serán grabadas, mientras que los exámenes serán presenciales (aula por definir).
- El curso es teórico-práctico y exige el uso frecuente de la computadora. Se espera no sólo que el estudiante aprenda los fundamentos teóricos de las técnicas multivariantes, sino que también aplique las técnicas a archivos de datos utilizando paquetes estadísticos.
- Presentaciones teóricas: se impartirán lecciones sincrónicas por parte del docente donde se explicarán los conceptos y sus aplicaciones. Las lecciones serán grabadas y estarán disponibles para que los estudiantes las puedan descargar.
- Prácticas: se realizarán laboratorios estructurados con ejercicios sobre los contenidos desarrollados en las clases teóricas. Estos laboratorios también se grabarán y la solución también estará disponible. Durante las sesiones de laboratorio se utilizará el lenguaje de programación R para realizar ejercicios de la materia vista en clase.
- Tareas: se asignarán ejercicios de práctica que incluirán aplicaciones con datos para ser analizados con R, así como interpretaciones de los resultados
- Trabajos de investigación: con el objetivo de poner en práctica los conocimientos, los estudiantes deberán enfrentarse a un problema real que deben analizar y presentar en forma de artículo.

6. Evaluación

- Se realizarán dos exámenes parciales, en ellos se evaluarán conceptos y la forma de interpretar resultados.
- Los estudiantes presentarán dos trabajos de análisis de datos reales. Los trabajos deberán presentarse en forma de artículos que sigan los lineamientos establecidos por la Revista Serengueti, con un máximo de extensión de 12 páginas.
- El primer trabajo debe incluir **técnicas de agrupamiento** y los estudiantes deberán diseñar un instrumento para recolectar los datos y luego hacer el análisis.
- El segundo trabajo puede incluir **técnicas de clasificación**, para lo cual podrán utilizar una base de datos existente.

Primer examen	25%
Segundo examen	25%
Primer artículo	20%
Segundo artículo	20%
Trabajo adicional	10%

7. Cronograma

	Módulo	K	J	Actividad
MAR	Análisis de componentes principales (PCA)	29		
			31	
ABRIL		5		
			7	
		12		SEMANA SANTA
			14	
		19		Presentación aplicaciones PCA
	Análisis de agrupamiento (clúster) y escalamiento multidimensional (MDS)		21	
		26		
MAYO			28	
		3		
			5	
		10		
			12	
		17		
			19	
	Técnicas de clasificación	24		
			26	
		31		Examen No.1
JUNIO			2	Artículo 1 (escrito y oral)
		7		
			9	
		14		
			16	
		21		
			23	
		28		
			30	
JULIO		5		
			7	
		12		Examen No.2
			14	
		19		
			21	Artículo 2 (escrito y oral)
		26		
			28	

8. Referencias bibliográficas

Cichosz, Pawel. (2015). Data Mining Algorithms: Explained Using R. Wiley.

Everitt, B y Hothorn, T. (2011). An Introduction to Applied Multivariate Analysis with R. Springer
BIBLIOTECA LUIS DEMETRIO TINOCO 519.535.028.5 E93i

Hair, J.F. et al (2014). Multivariate Data Analysis. Pearson Education Limited.
BIBLIOTECA LUIS DEMETRIO TINOCO 519.535 M958m7 2015

James, G., Witten, D., Hastie, T. y Tibshirani. R. (2021). An Introduction to Statistical Learning: with Applications in R (2da. Ed). Springer.

Johnson, R. A. y Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Prentice-Hall International, Inc.
BIBLIOTECA LUIS DEMETRIO TINOCO 519.535 J68a6

Mirkin, B (2005). Clustering for Data Mining: A Data Recovery Approach. Chapman & Hall.

Mishra, P. (2016). R Data Mining Blueprints. Packt Publishing.

Pan et al. (2013). Introduction to Data Mining. Pearson.

Olson et al. (2017). Predictive Data Mining Models. Springer.

Ramasubramanian, K y Singh, A (2017). Machine Learning Using R. Apress.

Sarkar, D (2008). Lattice: Multivariate Data Visualization with R. Springer.
BIBLIOTECA LUIS DEMETRIO TINOCO 006.6 S245L