UNIVERSIDAD DE COSTA RICA

EEs Escuela de Estadística

XS-4110: INTRODUCCION AL ANALISIS MULTIVARIADO

PROGRAMA

I SEMESTRE 2023

	GRUPO 01	**GRUPO 02**
Docente:	Ricardo Alvarado Barrantes.	Shirley Rojas Salazar.
Correo:	estad.ucr@gmail.com	ucrsrs@gmail.com
Teléfono:	84021263	87737058
Clases:	L-J: 1:00-2:50 pm	L-J: 1:00-2:50 pm
Consulta:	M: 4:00-6:00 pm	M: 9:00-11:00 am
Zoom:	https://udecr.zoom.us/j/86862167699	https://udecr.zoom.us/j/89709609762
Materiales:	https://www.dropbox.com/sh/gznbo3g5r152h1k/AAB1-KsP_uUhc2UgtIhcW1FEa?dl=0	

1. Descripción

En este curso se presentan diversas técnicas estadísticas multivariadas para minería de datos con un enfoque teórico-práctico. Se estudia el análisis de componentes principales, el análisis de agrupamiento, el escalamiento multidimensional, técnicas de clasificación y técnicas de clasificación basadas en árboles. Además de los conocimientos teóricos, el estudiante debe aplicar las técnicas usando lenguajes de programación en prácticas y 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 de minería de datos, gráficas y cuantitativas, para contribuir en la resolución de problemas que se presenten en distintos contextos mediante la interpretación adecuada de los resultados.

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 de los componentes principales. 1.2 Construcción: valores y vectores propios, variancia explicada, cálculo de los puntajes en los componentes principales, uso de covariancias o correlaciones, número de componentes principales. 1.3 Representación gráfica: biplot. 1.4 Evaluación de resultados: reproducción de matriz de variancias, correlación entre componentes y variables originales.
II. *Análisis de agrupamientos y escalamiento multidimensional (MDS)*
2.1 Distancias entre individuos (variables continuas, nominales, mezclas). 2.2 Distancias entre grupos (vecino más cercano, vecino más lejano, salto promedio). 2.3 Selección de variables para el análisis / estandarización. 2.4 Agrupamientos jerárquicos: algoritmo, representación (dendograma). 2.5 Métodos de k-medias y k-medoides: algoritmo. selección del número de clusters. 2.6 Validación: número de clusters. 2.7 Presentación de resultados: mapas de calor. 2.8 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 3.2 Métodos basados en árboles: a) Arboles de decisión b) Agregación de Bootstrap (bagging), bosques aleatorios (random forest) c) Boosting d) BART 3.3 Evaluación del modelo: a) Métricas de desempeño. b) Validación cruzada.

5. Metodología

El curso es teórico-práctico y exige el uso frecuente de la computadora. Se espera que el estudiante aprenda los fundamentos teóricos de las técnicas de análisis multivariado y que aplique las técnicas a archivos de datos utilizando lenguajes de programación estadística. Se propone una combinación de actividades, tales como:

1. Presentaciones teóricas: lecciones por parte del docente donde se explican los conceptos y sus aplicaciones.
2. Ejercicios en clase para que las sesiones sean activas.
3. Laboratorios: sesiones estructuradas con ejercicios sobre los contenidos desarrollados en las clases teóricas con solución disponible. Durante las sesiones de laboratorio se utiliza el lenguaje de programación R.
4. Prácticas: ejercicios fuera de clase que incluyen aplicaciones con datos para ser analizados, así como interpretaciones de los resultados.
5. Trabajos de investigación: con el objetivo de poner en práctica los conocimientos, el estudiantado debe enfrentar problemas reales que debe analizar y presentar con el formato de un artículo científico.
6. Revisión bibliográfica: los estudiantes deben seleccionar un artículo de una revista científica donde se aplique el PCA, exponerlo y entregar un reporte sobre el artículo con énfasis en la metodología, incluyendo una crítica sobre el mismo.

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%
Trabajos adicionales	10%

7. Cronograma

	Módulo	L	J	Actividad
MARZO	Análisis de componentes principales (PCA)	13		
			16	
		20		
			23	
	Análisis de agrupamiento (clúster) y escalamiento multidimensional (MDS)	27		
			30	
		3		SEMANA SANTA
			6	
		10		FERIADO
ABRIL			13	Presentación aplicaciones PCA
		17		
			20	
		24		
			27	
		1		FERIADO
			4	
	Técnicas de clasificación	8		Anteproyecto
			11	
MAYO		15		
			18	Examen No.1
		22		
			25	Artículo 1 (escrito y oral)
		29		
			1	
		5		
			8	
		12		
JUNIO			15	
		19		
			22	
		26		
			29	Examen No.2
		3		
			6	Artículo 2 (escrito y oral)
JULIO		10		
			13	

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