UNIVERSIDAD DE COSTA RICA

EEs Escuela de Estadística

XS-41100: INTRODUCCION AL ANALISIS MULTIVARIADO

PROGRAMA

I SEMESTRE 2020

Docente:	Ricardo Alvarado Barrantes.
Oficina:	17 Estadística
Teléfono	8402-1263
Correo electrónico:	estad.ucr@gmail.com
Horas de consulta:	V: 1:00-3:00
Horario de clases:	M: 1:00-4:50pm

1. Descripción

Curso introductorio de técnicas estadísticas multivariados 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.

- 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

Ofrecer una visión general de las técnicas básicas, gráficas y cuantitativas, del análisis multivariado que involucra varias variables y múltiples casos, e ilustrar sus aplicaciones con datos provenientes de nuestro medio y de revistas científicas.

3. Objetivos Específicos

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

- Resolver problemas con las técnicas estadísticas multivariados básicas: análisis de componentes principales, y análisis de agrupamientos.
- Resolver problemas con las técnicas estadísticas multivariados de clasificación bajo un enfoque general: análisis discriminante, regresión logística, árboles de decisión, k-vecinos más cercanos.
- Aplicar los métodos de ensamblaje de modelos a las técnicas de clasificación.
- Reconocer situaciones donde se puedan aplicar las técnicas aprendidas.
- Verificar los supuestos en que se apoyan las técnicas antes de aplicarlas.
- Evaluar las bondades y limitaciones de las técnicas.
- Procesar datos multivariados utilizando software estadístico, e interpretar los resultados obtenidos.

4. Contenidos

I. *Análisis de componentes principales (PCA) – 3 clases*
1.1 Objetivos del PCA 1.2 Características de los componentes principales 1.3 Construcción de los componentes principales 1.4 Uso de covariancias o correlaciones 1.5 Cálculo de los puntajes en los componentes principales 1.6 Variancia explicada 1.7 Representación gráfica: biplot 1.8 Número de componentes principales 1.9 Evaluación de resultados: a) Reproducción de matriz de variancias b) Correlación entre componentes y variables originales
II. *Análisis de agrupamientos (clústers) – 4 clases*
2.1 Objetivos del análisis de agrupamientos 2.2 Distancias entre individuos (variables continuas, nominales, mezclas) 2.3 Selección de variables para el análisis / Estandarización 2.4 Distancias entre grupos (vecino más cercano, vecino más lejano, salto promedio) 2.5 Agrupamientos jerárquicos: algoritmo y representación (dendograma) 2.6 Método de k-medias: algoritmo y selección del número de clústers 2.7 Validación: número de clusters. 2.8 Presentación de resultados: mapas de calor
III. *Clasificación – 5 clases*
3.1 Técnicas de clasificación: a) Análisis discriminante b) Regresión logística binomial y multinomial c) Arboles de decisión d) K-vecinos más cercanos e) Máquinas de vectores de soporte f) Redes neuronales 3.2 Validación 3.3 Métricas de desempeño
IV. *Ensambles de modelos – 2 clases*
4.1 Agregación de bootstrap 4.2 Bosques aleatorios 4.3 Boosting 4.4 Stacking

5. Metodología

- El curso seguirá la modalidad virtual usando la plataforma Zoom para las lecciones, las cuales serán grabadas.
- 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 tres 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**.
- El segundo trabajo puede incluir **técnicas de clasificación**.

Primer examen	20%
Segundo examen	25%
Tercer examen	25%
Primer artículo	15%
Segundo artículo	15%

7. Cronograma

	M	Actividad
ABR	7	PCA
	14	PCA
	21	PCA
	28	CLUSTER
MAY	**5**	**Examen No.1**
	12	CLUSTER
	19	CLUSTER
	26	CLUSTER
JUN	2	CLASIFICACION
	9	**Examen No.2**
	16	CLASIFICACION / **Artículo 1**
	23	CLASIFICACION
	30	CLASIFICACION
JUL	7	CLASIFICACION
	14	ENSAMBLE
	21	ENSAMBLE / **Artículo 2**
	28	**Examen No.3**

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

Hernández R, Óscar (1998). Temas de Análisis Estadístico Multivariado. Editorial UCR.
BIBLIOTECA LUIS DEMETRIO TINOCO 519.535 H557t

Hernández R, Óscar (2006). Notas adicionales a Temas de Análisis Estadístico Multivariado.

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

Johnson, D. (1998). Métodos multivariados aplicados al análisis de datos. International Thompson Editores.

Kleinbaum et al. (1998). Applied Regression Analysis and other Multivariate Methods. Duxbury Press.

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.

Pla, E.L. (1986). Análisis Multivariado: Método de componentes principales. O.E.A. Washington. (Cap 4: Caracterización de la producción lechera de un distrito).
BIBLIOTECA LUIS DEMETRIO TINOCO 519.535 P696a

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