



Universidad de Costa Rica
Sistema de Estudios de Posgrado
Programa de Posgrado en Estadística

NP-1602 Introducción al Análisis Multivariado

INFORMACIÓN GENERAL

Plan de estudios al que pertenece el curso:	730501 Maestría Académica en Estadística y 730506 Maestría Profesional en Estadística
Tipo de curso:	Nivelatorio para la Maestría Académica en Estadística y Maestría Profesional en Estadística
Modalidad:	Presencial- Bajo virtual Se utilizará la plataforma mediación Virtual
Número de créditos:	4 créditos
Horas presenciales:	4 horas semanales
Horario del curso:	Jueves 18:00-21:50
Aula:	300-A de <i>Ciencias económicas. Algunas clases en laboratorios.</i>
Horas de atención de estudiantes:	2 horas semanales (<i>Martes, 19:00 a 21:00</i>)
Requisitos:	No tiene
Correquisitos:	No tiene
Periodo lectivo:	I-2023
Profesor:	<i>Máster Deiby Solano Cambroner.</i>



PROGRAMA

Descripción y justificación del curso

Curso introductorio al análisis multivariado para estudiantes que ingresan a la Maestría Profesional o Académica en Estadística que provienen de disciplinas distintas a la Estadística o bien, que nunca hayan aprobado un curso similar previamente. Se pretende con este ofrecer el conocimiento mínimo necesario para nivelarse con los estudiantes que provienen de la carrera de Estadística. Con este curso, y el resto de cursos de nivelación, los estudiantes meta se integrarán exitosamente los cursos regulares del programa. Este curso está enfocado en el aprendizaje de los principales métodos de análisis para variables numéricas como no numéricas. Además, se centra en conocer la razón que justifica la aplicación de un determinado método, su objetivo, desarrollo e inmediata aplicación.

Se brindará al estudiante la posibilidad de ampliar los métodos mediante el uso del lenguaje de programación (R).

Objetivo general

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

Objetivos específicos

Al final del curso se espera que el estudiante haya aprendido a:

- 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 R , e interpretar los resultados obtenidos.



Contenidos del curso

- I. Análisis de componentes principales (ACP)
 - a. Objetivos del ACP
 - b. Características de los componentes principales
 - c. Construcción de los componentes principales
 - d. Uso de covariancias o correlaciones
 - e. Cálculo de los puntajes en los componentes principales
 - f. Variancia explicada
 - g. Representación gráfica: biplot
 - h. Número de componentes principales
 - i. Evaluación de resultados

- II. Análisis de agrupamientos
 - a. Objetivos del análisis de agrupamientos
 - b. Distancias entre individuos (variables continuas, nominales, mezclas)
 - c. Selección de variables para el análisis / Estandarización
 - d. Distancias entre grupos (vecino más cercano, vecino más lejano, salto promedio)
 - e. Agrupamientos jerárquicos: algoritmo y representación (dendograma)
 - f. Método de k-medias: algoritmo y selección del número de clústers
 - g. Presentación de resultados

- III. Clasificación
 - a. Técnicas de clasificación:
 - i. Análisis discriminante
 - ii. Regresión logística binomial y multinomial
 - iii. Árboles de decisión
 - iv. K-vecinos más cercanos
 - v. Máquinas de vectores de soporte
 - vi. Redes neuronales
 - b. Validación
 - c. Métricas de desempeño

- IV. Ensamblados de modelos
 - a. Agregación de bootstrap
 - b. Bosques aleatorios
 - c. Boosting
 - d. Stacking

Metodología

El curso es teórico-práctico y donde se realizarán presentaciones teóricas, se impartirán lecciones magistrales por parte del docente donde se explicarán los conceptos y sus aplicaciones. Además, se



realizarán laboratorios en computadora que se desarrollaran mediante el programa R durante el semestre. Se resolverán los laboratorios y se brindarán algunas tareas evaluadas como práctica.

Los estudiantes realizarán trabajos de análisis de datos reales, donde las aplicaciones podrán ser tomadas de investigaciones en áreas como la economía, medicina, psicología, biología, población, etc.

Los estudiantes deberán escribir un documento con formato de artículo científico con las características y en las fechas brindadas por el profesor. Los resultados deberán ser expuestos en forma oral en una presentación con tiempo limitado. La asistencia a las lecciones será deseable, además, en las mismas se brindará materia para complementar los textos y las diapositivas que también será evaluada en exámenes o tareas, este material estará disponible en la plataforma de mediación virtual.

Evaluación

Se realizarán dos exámenes parciales, en ellos se evaluarán los conceptos, el uso del software para obtener resultados y la forma de interpretar resultados. Además, los estudiantes realizaran presentaciones y/o artículos sobre las técnicas vistas en el curso y tareas ya sea escritas y/o con presentación en clase.

Para la evaluación del trabajo de los artículos y presentaciones se proporcionará un desglose de la distribución de criterios, en los cuales se tomará en cuenta la revisión bibliográfica, formato de cuadros y gráficos estadísticos, la redacción, el planteamiento, y la profundidad en el análisis estadístico.

Si un estudiante faltase a algún examen o a la entrega de alguna tarea o artículo por causa justificada, debe solicitar por escrito la reposición indicando las razones de la ausencia, acompañada de los documentos justificantes. La misma debe entregarse ante el profesor a más tardar en cinco días hábiles del reintegro a lecciones. Entre los entornos a utilizar en el curso, el mismo será presencial bajo virtual: con mezcla de utilización de la plataforma mediación virtual, el correo electrónico y en algunas ocasiones, se podría requerir la plataforma Zoom. Las evaluaciones se utilizará una o una mezcla de las plataformas mencionadas anteriormente.

El cronograma de exámenes y la evaluación del curso se desglosa según los siguientes rubros:

Primer examen parcial (Cap. I y II, aprox. semana 7-8)	30%
Segundo examen parcial (Cap. III y IV, aprox. semana 16-17)	30%
Tarea (as)	10%
Artículos y presentaciones sobre técnicas	30%
Total	100%

Cronograma

Se indica entre paréntesis el número estimado de semanas para cada capítulo:

I(3), II(3), III(5), IV(3)



Bibliografía

- Hernández R, Óscar (1998). Temas de Análisis Estadístico Multivariado. Editorial UCR.
BIBLIOTECA LUIS DEMETRIO TINOCO 519.535 H557t
- Everitt, B y Hothorn, T. (2011). An Introduction to Applied Multivariate Analysis with R. Springer
BIBLIOTECA LUIS DEMETRIO TINOCO 519.535.028.5 E93i
- Zelterman Daniel. (2015). Applied Multivariate statistic with R. Springer International Publishing
- Matloff, Norman (2011). The art of R programming
- Santana, Julio y Farfan, Efraín (2014). El arte de programar en R, un lenguaje para la estadística. México : Instituto Mexicano de Tecnología del Agua. UNESCO. Comité Nacional Mexicano del Programa Hidrológico Internacional, ©2014.
- Hair, J.F. et al (2014). Multivariate Data Analysis. Pearson Education Limited.
BIBLIOTECA LUIS DEMETRIO TINOCO 519.535 M958m7 2015
- 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.
- Mirkin, B (2005). Clustering for Data Mining: A Data Recovery Approach. Chapman & Hall.
- 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.
- Sarkar, D (2008). Lattice: Multivariate Data Visualization with R. Springer.
BIBLIOTECA LUIS DEMETRIO TINOCO 006.6 S245L