



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:	Teórico
Número de créditos:	4 créditos
Horas presenciales:	4 horas semanales - M: 17:00 - 20:50
Horas de atención de estudiantes:	1 horas semanales - V: 18:00 - 18:50
Requisitos:	No tiene
Correquisitos:	No tiene
Profesor:	Oscar Centeno Mora



PROGRAMA

Descripción general del curso

El curso de Introducción al Análisis Multivariado está enfocado en el aprendizaje de los principales métodos de análisis para datos cuantitativos, cualitativos, texto y visualización de la data. El curso se centra en conocer cierto método multivariado, su objetivo analítico, el desarrollo e inmediata aplicación. Adicional, se enfatiza en las técnicas de visualización de la data tanto en los métodos multivariados como en la implementación de dashboards.

Justificación

El presente curso introduce los fundamentos del análisis multivariado para estudiantes que recién están ingresando a la Maestría Profesional o Académica en Estadística, y que además provienen de disciplinas distintas a la Estadística, o bien, que nunca hayan aprobado un curso similar previamente. El fin último es nivelar y producir un ingreso exitoso a los estudiantes que tiene por objetivo llevar los cursos del programa Maestría Profesional o Académica en Estadística.

Objetivo general

Ofrecer una visión general de las técnicas multivariadas, además de la consumación gráfica en datos cuantitativos, cualitativos, textos, y otros, así como la implementación de la visualización de datos mediando dashboards.

Objetivos específicos

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

1. Resolver problemas con las técnicas estadísticas multivariantes básicas: visualización de datos, técnicas multivariadas cuantitativas, técnicas multivariadas textuales y de conexiones, y técnicas multivariadas para datos mixtos.
2. Reconocer situaciones donde se puedan aplicar las técnicas aprendidas.
3. Formulación de dashboards bajo diferentes contextos analíticos.
4. Evaluar las bondades y limitaciones de las técnicas multivariadas.
5. Presentar los resultados de todas las técnicas multivariadas además de los dashboards.
6. Aplicar un trabajo final en relación con una de las técnicas estudiadas en el curso, conjuntamente con la formulación de un dashboard.
7. La utilización de los paquetes de análisis de la información: R Studio, shinydashboard, Excel y Tableau.



Metodología

El curso es teórico-práctico y exige el uso frecuente de la computadora. Se espera no solo que el estudiante aprenda los fundamentos teóricos de las técnicas multivariantes, sino que también aplique las técnicas a archivos de datos utilizando paquetes estadísticos. Las lecciones son magistrales, con ilustraciones prácticas de cada una de las técnicas multivariadas o de la formulación de los dashboards bajo diferentes contextos, para luego ser acompañadas por un laboratorio. Adicional, se impartirán lecturas de artículos además de referencias de páginas web para familiarizar a los estudiantes con aplicaciones prácticas de las técnicas multivariantes o de las representaciones gráficas y / o dashboards. Al ser un curso interactivo entre alumno-profesor, se solicita la participación de las y los estudiante para mejorar la calidad del curso y del aprendizaje conjunto de los participantes.

Presencia obligatoria

Se recuerda la presencia obligatoria a todas las clases. Si por razones mayores el estudiante no puede asistir, es imperativo justificar la ausencia. Además, la no presentación de cualquier trabajo u otro similar que posea un porcentaje de la calificación del curso, será reportado como porcentualmente nulo.

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. Los exámenes se realizarán en computadora, y los estudiantes deberán conocer el manejo del software para obtener los resultados. Por otra parte, los estudiantes presentarán varios trabajos de análisis de datos reales. Los trabajos deberán presentarse en forma de artículos cortos. Si un estudiante faltase a algún examen por causa justificada, debe solicitar por escrito la reposición del examen indicando las razones de la ausencia, acompañada de los documentos justificantes. La misma debe entregarse ante el profesor que imparte el curso a más tardar en cinco días hábiles del reintegro a lecciones. La evaluación del curso se desglosa según los siguientes rubros:

Rubro	Porcentaje (%)
Parcial 1	15
Parcial 2	15
Presentaciones e informes de las técnicas multivariadas ¹	40
Presentación del Story line y de los dashboards ²	20
Trabajo Final	10
TOTAL	100

¹ Constituido de 8 trabajos, cada uno con un valor de 5%

² Constituido de 4 trabajos, cada uno con un valor de 5%



Descripción del curso

TEMAS Y CONTENIDOS	MATERIALES
Tema 1 Visualización de datos univariados y multivariados: construyendo un Story-line Principales representaciones unidimensionales Representación multidimensional: matrices de gráficos, colores, gráficos tridimensionales, contornos, etc. Representación de correlaciones: correlograma Detección de valores extremos multivariados: boxplot bivariado, envoltorio convexo, chi-plot Mapas y otras representaciones gráficas Construcción de un Story-line	PC
Tema 2 Análisis por Componentes Principales (PCA) Elementos de álgebra lineal Objetivos del PCA Características de los componentes principales Construcción de los componentes principales Uso de covariancias o correlaciones Cálculo de los puntajes en los componentes principales Variancia explicada Representación gráfica: biplot Número de componentes principales Evaluación de resultados: 1) reproducción de matriz de variancias, 2) correlación entre componentes y variables originales. 3)	PC
Tema 3 Análisis Factorial (FA) Diferencia entre el PCA y el FA Modelos de factores comunes, modelos ortogonales Métodos de estimación: 1) componentes principales, 2) máxima verosimilitud Selección del número de factores Rotación de factores Estimación de puntajes factoriales Pensamiento a priori: el análisis factorial confirmatorio	PC
Tema 4 Los Dashboards (parte 1) El dashboard en Excel	



Tema 5 Análisis discriminante Separación y clasificación Regla basada en la minimización del costo esperado de clasificación errónea Procedimiento para poblaciones multinormales con matrices de covariancias iguales Evaluación de la eficacia de las reglas de clasificación Función discriminante de Fisher Tabla de clasificación.	PC
Tema 6 Los Dashboards (parte 2) El dashboard en Tableau	PC
Parcial #1 Temas 1 - 6	
Tema 7 Análisis de agrupamiento (clusters) Objetivo del análisis de agrupamiento Selección de variables para el análisis Similitud entre objetos: 1) distancias para variables continuas, 2) distancias para variables binarias, 3) distancia para mezcla de variables. Estandarización Distancias entre grupos Agrupamientos jerárquicos: algoritmo y representación (dendograma) Método de k-medias: algoritmo y selección del número de clústers. Presentación de resultados: 1) componentes principales, 2) sombras y vecindarios, 3) rayas, 4) perfiles Clusters para variables categóricas Relación entre PCA-Discriminación-Agrupamiento Objetivo de los métodos Supuestos y comportamiento de los datos La reducción, clasificación y la discriminación	PC
Tema 8 Los Dashboards (parte 3) El dashboard en Shinydashboard	PC
Tema 9 Análisis de texto La relación de las palabras Cómo hacer la frecuencia de palabras Análisis de sentimiento La nube de palabra	



Tema 10 Análisis de redes ¿Qué es una red? Por qué analizar de unidades por sus relaciones Centro, relación fuerte, baja y otros conceptos de las redes. Similaridades y unidades de influencia La red social y su aplicación	
Tema 11 Análisis de afinidad: Market Basquet Analysis Cómo seleccionar las cosas ¿Frecuencias, objetivos o qué? Discriminación las canastas Determinar la mejor afinidad por un grupo de objetos El Market Basquet Análisis y su aplicación	
Tema 12 Árboles de decisión y de regresión El análisis supervisado vs el análisis descriptivo Árboles de decisión y de regresión ¿Por qué estudiar los árboles? Clasificación y predicción Reglas de partición El método CART y CHAID Evaluación del árbol: nodos y profundidad Los árboles y su aplicación Estimación de los árboles Descripción del árbol	PC
Parcial # 2 Temas 7 - 12	PC



Cronograma

Mes	Tema	Día	Actividad
Marzo	Presentación del curso Introducción del curso R Studio - Markdown Tema 1	13	
	Tema 2	20	
	Tema 3	27	Presentación T2
Abril	Tema 4	3	Presentación T3
	Tema 5	10	Presentación T1
		17	
	Tema 6	24	Presentación Tema 5
Mayo		1	
	Examen parcial n°1	8	
	Tema 7	15	Presentación Tema 4
	Tema 8	22	Presentación Tema 7
	Tema 9	29	Presentación Tema 6
Junio	Tema 10	5	Presentación Tema 9
	Tema 11	12	Presentación Tema 10
	Tema 12	19	Presentación Tema 11
			Presentación Tema 8
	Trabajo final	26	Presentación Tema 12
			Presentación trabajo final
Julio	Examen parcial n°2	3	
	Entrega de resultados	10	



Bibliografía

La siguiente bibliografía es la recomendada para el curso:

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