

Planificación de la Asignatura: Exploración de Datos Multivariados

Fecha: 23/10/2024 13:02

Código: T1634

Carrera: Tecnicatura Universitaria en Procesamiento y Explotación de Datos

Departamento Académico: Matemática

Docente a cargo:

Correo del docente a cargo: sin datos

Régimen de Dictado: Cuatrimestral 1º Cuatrimestre

Carga Horaria Semanal: 5 horas semanales

Carga Horaria Total: 70 horas

Contenidos Mínimos:

Análisis exploratorio para el caso multivariado de datos cualitativos y cuantitativos. Técnicas descriptivas multidimensionales. Análisis en componentes principales. Análisis factorial. Análisis de correspondencias.

Competencias Genéricas:

No corresponde.

Competencias Específicas:

No corresponde.

Argumentación de aportes marcados en la matriz de competencias:

Correlativas Regulares para cursar:

Comprensión Lectora y Producción Escrita,
Espacio Integrador I,
Probabilidad y Estadística;
Algoritmos y Estructura de Datos
Aspectos Legales del Uso de la Información

Correlativas Aprobadas para cursar:

Informática Básica
Algebra y Calculo

Correlativas Aprobadas para promocionar o rendir el examen final:

No posee

Insercion de la Asignatura en el plan de Estudios:

La asignatura cobra una importancia fundamental en el perfil del egresado, ya que provee las herramientas esenciales para describir, organizar, procesar y explorar datos de diversas naturalezas, posibilitando la extracción de información valiosa. Esta materia establece una conexión intrínseca con los contenidos de asignaturas del primer año, como Álgebra, Cálculo, y Probabilidad y Estadística. Además, se erige como un pilar esencial para cátedras subsiguientes como Modelado Estadístico, Aprendizaje Maquinal, Minería de Datos, Visualización de la Información y, por supuesto, con los Espacios Integradores.

La exploración de datos, que constituye el primer paso en la aplicación de técnicas de minería de datos y aprendizaje automático, resulta crucial. Sin este proceso, se torna imposible comprender el problema en su totalidad, analizar las relaciones entre variables y tomar decisiones que puedan mejorar el rendimiento de algoritmos tanto en clasificación como en agrupación. Asimismo, el análisis multivariado aplicado a conjuntos masivos de datos posibilita un ahorro significativo en tiempo y costo computacional.

En resumen, esta asignatura no solo es un fundamento para la comprensión profunda de otras materias, sino que también desempeña un papel central en el desarrollo de habilidades esenciales para abordar problemas complejos, tomar decisiones informadas y optimizar procesos computacionales en el ámbito de la minería de datos y el aprendizaje automático.

Objetivo General:

Conocer las alternativas y estrategias disponibles para el análisis, representación y resumen de datos multivariados tanto cuantitativos como cualitativos, acentuando la necesidad de la evaluación comparativa de diferentes técnicas disponibles para un mismo problema, e interpretar críticamente los resultados obtenidos.

Objetivos Particulares:

- Conocer las principales técnicas del análisis estadístico de varias variables (multivariado).
- Aplicar técnicas de estadística multivariada dependiendo del fenómeno a tratar y que facilitan los análisis y conclusiones de las bases de datos.
- Utilizar métodos de análisis multivariantes para aclarar y explicar las relaciones entre las diferentes variables que pueden estar asociadas con la recopilación y evaluación de datos estadísticos.
- Conocer las reglas básicas para visualizar datos multivariados
- Desarrollar un sentido crítico para analizar y comunicar datos; y para responder preguntas basadas en la visualización de los datos.
- Aprender la estructura estadística de los métodos de reducción de dimensionalidad, así como su aplicación para problemas de compresión, visualización y estimación de datos faltantes.
- Desarrollar hábitos y formas de estudio propios del nivel universitario.

Programa Analítico:

Los contenidos mínimos de la asignatura se proponen encuadrados en seis unidades temáticas conformando el siguiente programa analítico:

UNIDAD 1: Conceptos básicos. Variables estadísticas y datos estadísticos. Clasificación de las variables de acuerdo a su naturaleza. Recolección y organización de datos. Análisis de datos univariado. Medidas de posición: media, mediana, modo, cuartiles. Medidas de variabilidad o dispersión: rango, rango intercuartil, desvío estándar, coeficiente de variación. Medidas para caracterizar la distribución: coeficiente de simetría y coeficiente de curtosis. Representación gráfica acorde a la naturaleza de la variable. Reconocimiento de datos atípicos y faltantes. Comparación de datos. Análisis e interpretación de la información del análisis univariado.

UNIDAD 2: Análisis bivariado. Datos bivariados. Tabla cruzada de datos categóricos. Gráfico de mosaico. Análisis bivariado de datos cuantitativos. Diagrama de dispersión. Análisis e interpretación para detectar puntos anómalos, y para extraer información multivariada. Comparación de datos. Transformaciones del conjunto de datos.

UNIDAD 3: Análisis Multivariado. Medidas de posición y dispersión de datos multivariados. Matriz de Covarianza y correlación. Criterios para usar matrices de covarianza o de correlación. Dispersograma. Correlograma. Distancia de Mahalanobis. Detección de outlier multivariados. Medidas robustas para posición y escala. Análisis e interpretación de gráficos multivariados: gráficos tridimensionales, gráfico de coordenadas paralelas, gráfico de perfiles multivariados, gráfico de estrellas, gráfico radial, gráfico caras de Chernoff.

UNIDAD 4: Análisis de Componentes Principales. Objetivos e importancia del análisis de componentes principales. Reducción de la dimensión. Variabilidad explicada. Criterios para decidir la cantidad de componentes principales. Obtención de las coordenadas de los individuos. Obtención de las coordenadas de las variables. Información que aportan las cargas. Interpretación de las componentes principales. Representación gráfica biplot. Análisis e Interpretación de la información.

UNIDAD 5: Análisis Factorial de Correspondencias Simples y Múltiples. Contrastes de Homogeneidad e independencia. Test Chi Cuadrado. Objetivo e importancia del análisis de correspondencia. Análisis de correspondencia simple. Perfiles fila y columna. Inercia total.

Obtención de las coordenadas factoriales. Construcción e interpretación del Biplot simétrico.

Interpretación geométrica de la inercia. Análisis de correspondencia múltiple. Matriz disyuntiva. Matriz de Burt. Representación gráfica simultánea. Análisis e interpretación de las contribuciones por individuo y variable.

UNIDAD 6: Comparación de medias. Comparación de medias caso univariado. Test paramétricos y no paramétricos entre dos grupos. Test paramétricos y no paramétricos entre de tres o más grupos: análisis de la varianza de un factor. Comparación de medias caso multivariado test de Hotelling.

Metodología Didáctica:

La propuesta pedagógica se centra en clases teóricas con un enfoque eminentemente práctico. Al iniciar cada tema, nos sumergiremos en un problema de estudio específico que actuará como hilo conductor. A través de preguntas guía relacionadas con un dataset particular, discutiremos los métodos y los elementos teóricos esenciales, sin adentrarnos excesivamente en los fundamentos matemáticos. Este enfoque tiene como objetivo principal proporcionar a los estudiantes una comprensión clara de la situación y capacitarlos para extraer información relevante.

En la medida de lo posible, utilizaremos datasets que permitan establecer conexiones entre diferentes temas, haciendo hincapié especialmente en la interpretación de resultados y en la selección estratégica de la información extraída para su posterior comunicación.

El entorno de trabajo elegido será Jupyter, una plataforma que facilita la escritura y ejecución de código Python de manera accesible para todos los participantes. Además, Jupyter permite la integración del lenguaje R en el mismo archivo, junto con la capacidad de agregar líneas de texto explicativas. De esta manera, cada estudiante contará con un archivo compartido que reflejará el código utilizado en clase, enriquecido con las contribuciones individuales, así como anotaciones relevantes y análisis de la información obtenida.

Esta metodología busca fomentar la participación activa de los estudiantes, brindándoles una experiencia de aprendizaje práctica y colaborativa. Considero que esta aproximación dinámica no solo facilita la comprensión de los conceptos, sino que también promueve un ambiente propicio para el intercambio de ideas y el desarrollo conjunto de habilidades analíticas.

Formación Práctica:

Las clases prácticas constituirán una inmersión en el análisis de un dataset real, abordando su complejidad con la profundidad que exige. Las guías proporcionarán preguntas estratégicas que orientarán la búsqueda de respuestas y fomentarán la generación de interrogantes para análisis adicionales. Durante estas sesiones prácticas, se hará especial énfasis en la transposición de la teoría a la práctica, abordando las dificultades inherentes al análisis de datos.

Se explorará la necesidad del pensamiento crítico, la habilidad para enfrentar y superar obstáculos, así como la eficacia en la comunicación de los resultados obtenidos. Además, se destacará la importancia de discernir entre los resultados que aportan información valiosa al análisis y aquellos que no resultan relevantes.

Estas clases prácticas no solo buscan ofrecer a los estudiantes una experiencia inmersiva en el mundo del análisis de datos, sino también cultivar habilidades esenciales como el pensamiento analítico, la resolución de problemas y la capacidad de comunicar de manera efectiva los hallazgos. El enfoque estará en aprender de los desafíos prácticos, fomentando así un ambiente educativo que promueva el desarrollo integral de las habilidades críticas para el análisis de datos en contextos reales.

Listado de Actividades de Formación Práctica:

- Trabajo Práctico N°1: Limpieza y Manipulación de Datos.
- Trabajo Práctico N°2: Medidas de Resumen de Datos Univariados.
- Trabajo Práctico N°3: Gráficos para el Análisis de Datos Univariado.
- Trabajo Práctico N°4: Análisis de Datos Bivariado.
- Trabajo Práctico N°5: Comparación de dos Medias y Tipificación de Valores.
- Trabajo Práctico N°6: Comparación de más de dos medias.
- Trabajo Práctico N°7: Análisis Bivariado Cualitativo.
- Trabajo Práctico N°8: Análisis Multivariado Cuantitativo.
- Trabajo Práctico N°9: Reducción de Dimensionalidad.
- Trabajo Práctico N°10: Análisis Factorial de Correspondencias Simples y Múltiples.

Intensidad de la formación práctica

Detalle de la carga horaria total prevista para cada una de las siguientes actividades:

Actividades prácticas que aportan a las competencias específicas en el Nivel de dominio 1: 3 horas

Actividades prácticas que aportan a las competencias específicas en el Nivel de dominio 2: 3 horas

Actividades prácticas que aportan a las competencias específicas en el Nivel de dominio 3: 3 horas

Horas totales de actividades de formación práctica: 9 horas

Metodología de Evaluación Durante el cursado:

En sintonía con la metodología educativa de esta cátedra, se busca que la evaluación no solo sea un proceso de calificación, sino una oportunidad adicional de aprendizaje que contribuya al desarrollo profesional de los estudiantes. El perfil del egresado que aspiramos formar demanda habilidades como el trabajo colaborativo, la capacidad de abordar extensas bases de datos sin directrices precisas, y la destreza para realizar análisis en un plazo definido, seleccionando y comunicando de manera clara y precisa los resultados obtenidos. Para lograr esto, se organizarán equipos compuestos por dos integrantes desde la primera semana del curso, a los cuales se les asignará un conjunto de datos para su estudio. El proyecto se divide en dos partes, ambas requieren un análisis detallado utilizando las técnicas multivariadas impartidas durante las clases. Las guías de trabajo proporcionarán orientación para las tareas a realizar.

La primera parte del proyecto deberá ser presentada en la semana 8 del curso. Durante esta exposición, los grupos compartirán su trabajo con la profesora y sus compañeros, respondiendo preguntas basadas en su análisis. La segunda parte consistirá en ampliar y mejorar el proyecto inicial, aplicando todas las técnicas aprendidas durante el curso.

En ambas defensas, se deberá entregar un documento escrito, un archivo de Jupyter ejecutable, y la presentación en PowerPoint utilizada durante la exposición. El código empleado debe contener una descripción detallada de la base de datos, un análisis exploratorio y descriptivo de los datos, así como conclusiones contextualizadas dentro de los datos analizados.

Cabe destacar que cada grupo recibirá un conjunto de datos único, promoviendo así enfoques, análisis y desafíos distintos, fomentando un intercambio de experiencias en la presentación conjunta de los aprendizajes.

Esta metodología tiene como objetivo no solo evaluar conocimientos, sino también cultivar habilidades esenciales para el perfil profesional que se busca desarrollar en los estudiantes.

Metodología de Evaluación en Exámenes Finales:

- Los estudiantes que sean regulares deberán defender sus trabajos finales de forma individual y responder las preguntas del tribunal evaluador.
- Los estudiantes libres deberán solicitar un dataset a la cátedra diez días antes de la mesa, y durante la mesa deberán presentar un informe individual sobre el análisis multivariado del dataset en cuestión, y defenderlo ante el tribunal evaluador.

Condiciones de Regularidad :

Para Promocionar la materia se requiere:

- Asistir el 80% de las clases teóricas prácticas.
- Aprobar las dos instancias de defensa del TP integrador con calificación mayor o igual a 70% .

Para Regularizar la materia se requiere:

- Asistir el 70% de las clases teóricas prácticas. Aprobar las dos instancias de defensa del TP integrador con nota entre 50% y 69%.



Cronograma de parciales durante el primer Cuatrimestre:

Primer Examen Parcial: 22 de Abril de 2024

Segundo Examen Parcial: 10 de Junio de 2024

Tercer Examen Parcial: 19 de Junio de 2024

Cronograma de parciales durante el segundo Cuatrimestre:

Bibliografía Principal:

- Castaño. (2018). Análisis de datos multivariados XII SEMINARIO DE ESTADÍSTICA APLICADA III ESCUELA DE VERANO VII COLOQUIO REGIONAL DE ESTADÍSTICA INTRODUCCIÓN AL ANALISIS. Universidad Nacional de Colombia.
<https://www.studocu.com/pe/document/universidad-nacional-mayor-de-san-marcos/estadistica/casta-no-analisis-de-datos-multivariados/13585503>
- Chan, D., Badano, C., & Rey, A. (2019). Análisis inteligente de datos con lenguaje R: Con aplicaciones a imágenes. edUTecNe.
<https://1library.co/document/z3n15v7q-analisis-inteligente-datos-lenguaje-r-aplicaciones-imagenes.html>
- Díaz Monroy, L. G. (2007). Estadística multivariada: Inferencia y métodos (2. ed). Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia.
- García, J., Molina, J. M., Berlanga, A., Patricio, M. A., Bustamante, Á. L., & Padilla, W. R. (2018). Ciencia de datos. Técnicas analíticas y aprendizaje estadístico. Bogotá: Alfaomega Colombiana S.A.
- Joyanes, L. (Ed.). (2013.). Big Data: Análisis de grandes volúmenes de datos en organizaciones (Primera ed.). Alfaomega Grupo Editor, S.A. de C.V., México.
- Palacio, F., Apodaca, M., & Crisci, J. (2020). ANÁLISIS MULTIVARIADO PARA DATOS BIOLÓGICOS. Teoría y su aplicación utilizando el lenguaje R. VAZQUEZ MAZZINI EDITORES.
- Pardo, C. (2020). Estadística descriptiva multivariada. Universidad Nacional de Colombia.
<https://doi.org/10.36385/FCBOG-5-0>
- Peña, D. (2002). Análisis de Datos Multivariantes. University Carlos III de Madrid.
- Yaque, P. M. (2013). ANÁLISIS EXPLORATORIO DE DATOS CON R Y MINITAB. Universidad Complutense de Madrid.
- Posada Hernández, G. J. (2016). Elementos básicos de estadística descriptiva para el análisis de datos [Recurso electrónico]. Medellín: Funlam.

Bibliografía Complementaria:

Equipo de Cátedra:

Mg. Melisa Fernández

Actividades de Investigación Gestión y Extensión:

Requisitos de admisión para alumnos oyentes:

Infraestructura, equipamiento y recursos necesarios:

Otros: