

Universidad de Puerto Rico
Recínto de Río Piedras
Facultad de Administración de Empresas¹
Instituto de Estadística² y Sistemas Computarizados de Información
Bachillerato En Administración de Empresas

- I. **Título:** Minería de Datos
- II. **Codificación:** ESTA 5504
- III. **Número de Horas-Créditos:** 3 créditos, 3 horas semanales de conferencia y práctica
- IV. **Prerrequisito:** ESTA 3042 Estadística para la Administración de Empresas II. Estudiantes Graduados deberán contar con el permiso del Director.
- V. **Descripción:** Seminario de los conceptos fundamentales de minería de datos. Mediante el trabajo investigativo individual y colaborativo, se estudian las técnicas de aprendizaje automático que apoyan a la toma de decisiones al facilitar el análisis de grandes cantidades de datos. El contenido incluye técnicas de pre-procesamiento y visualización de datos, estudio y aplicación de técnicas de clasificación supervisada, clasificación no supervisada y métodos de detección de reglas de asociación. Se utilizan programas de aplicación que permiten al estudiante aplicar las técnicas estudiadas y el análisis de los resultados obtenidos. El estudiante preparará y presentará en forma escrita y oral, un proyecto de investigación donde aplique e integre conceptos del curso a un problema concreto.

VI. **Objetivos del Curso:**

Al finalizar el curso, el estudiante será capaz de

- 1) Describir conceptos y técnicas fundamentales en minería de datos.
- 2) Producir gráficas apropiadas en las fases de exploración y resultados de una tarea de minería de datos.
- 3) Aplicar técnicas de pre-procesamiento de datos, tales como tratamiento de datos faltantes, normalización, discretización, selección y extracción de variables.
- 4) Aplicar técnicas para crear modelos predictivos por medio de técnicas de clasificación supervisada.

¹ **Misión de la Facultad de Administración de Empresas:** Desarrollar líderes gerenciales, empresariales y académicos, profesionales y éticos, mediante una educación de excelencia e iniciativas de investigación y servicio en el contexto de Puerto Rico y el mundo.

² **Misión del programa de BAE en Estadística Aplicada:** El programa de Estadística Aplicada provee formación interdisciplinaria en los principios, metodologías y enfoques de la Estadística Aplicada, su fase computacional y la aplicación en escenarios diversos, en particular en la administración de empresas. De esta forma el programa contribuye a capacitar la sociedad en los enfoques analíticos para la investigación y la toma de decisiones con el fin de mejorar la calidad de vida de sus miembros.

- 5) Evaluar el rendimiento de los clasificadores, explicar .
- 6) Usar técnicas para la detección de reglas de asociación para descubrir hechos que ocurren en común dentro de un conjunto de datos para aplicarlas, por ejemplo, en los sistemas de recomendación (recommender systems)
- 7) Seleccionar el método de minería de datos adecuado para problemas específicos e información disponible.
- 8) Usar programas de aplicación para minería de datos y evaluar los resultados obtenidos.
- 9) Obtener datos de redes sociales siguiendo estándares éticos y sociales.
- 10) Usar software para representar, generar visualizaciones y obtener resultados de aplicación de los métodos de predicción y clasificación usando datos de redes sociales así como analizar e interpretar los resultados obtenidos.
- 11) Usar fuentes de datos textuales tales como documentos, correos electrónicos, contenido de blogs y redes sociales para usarlas en tareas de clasificación supervisada.
- 12) Comunicar los resultados obtenidos correctamente y de una manera clara y organizada, escrita y verbal.
- 13) Identificar implicaciones éticas y sociales del uso de minería de datos.
- 14) Demostrar una actitud crítica hacia la aplicación de métodos de minería de datos en la solución de una diversidad de problemas.

VII. Bosquejo del contenido y distribución del tiempo.

| <i>Temas</i> | <i>Distribución de tiempo (horas)</i> |
|---|---------------------------------------|
| I. Introducción <ol style="list-style-type: none"> a. ¿Qué es minería de datos? b. Aplicaciones de minería de datos. c. Pasos en el proceso de minería de datos. d. Consideraciones éticas en minería de datos. | 1.5 |
| II. Pre-procesamiento de datos <ol style="list-style-type: none"> a. Tratamiento de datos faltantes b. Reducción de la dimensionalidad: <ol style="list-style-type: none"> i. Selección de variables ii. Extracción de variables c. Discretización d. Normalización | 9 |
| III. Visualización <ol style="list-style-type: none"> a. Graficas de barras, scatterplots, boxplots e histogramas. b. "Heatmaps": Visualizar correlación y valores perdidos a. Visualización multidimensional: uso de color, tamaño y formas para representar más variables en una gráfica, uso de paneles múltiples y gráficas de coordenadas paralelas b. Otras gráficas especializadas: Visualización de datos en redes, datos jerárquicos y datos geográficos | 6 |

| | |
|---|-----------|
| | |
| IV. Clasificación Supervisada a. Clasificación Bayesiana b. Evaluación de un clasificador c. Clasificación usando vecinos cercanos (KNN) d. Árboles de decisión e. Otros métodos de clasificación: Neural Networks, Support Vector Machines | 9 |
| VI. Reglas de asociación a. Análisis de la canasta de Mercado b. El algoritmo a priori c. De conjuntos de elementos (“itemsets”) frecuentes a reglas de asociación d. Minería de reglas de asociación multidimensional e. De reglas de asociación a análisis de correlación | 6 |
| V. Analítica de Datos a. Analítica de Redes Sociales: Obtener datos de redes sociales, visualización y análisis de datos de redes sociales, métricas de datos sociales y su uso en predicción y clasificación. b. Text Mining: Preprocesamiento de documentos de texto, representación de texto en matrices, aplicación de métodos de minería de datos a datos textuales. | 9 |
| Exámenes y presentaciones | 4.5 |
| Total | 45 |

VIII. **Estrategias Instruccionales:** Conferencias, ejercicios de aplicación en clase, asignaciones y discusión de ejercicios. Uso del programa R a través de todo el curso y otros programas para aplicaciones para temas específicos como Tableau para visualización, SQL Server u APEX oracle para usar SQL y otros. Los estudiantes realizarán trabajos individuales y realizarán un proyecto grupal para el cual deberán realizar una presentación oral y preparar un informe final.

Este es un curso presencial, acorde a la Certificación 112 del 2014-2015 de la Junta de Gobierno de la UPR, un curso presencial es un curso en el cual un 75% o más de las horas de instrucción requieren la presencia física del estudiante y el profesor en el salón de clase. Esta definición permite que, de ser necesario, el 25% de las horas contacto puede ofrecerse utilizando otra modalidad.

IX. **Recursos de aprendizaje:** Uso del programa R o algún otro programa para minería de datos tal como RapidMiner u Orange. El estudiante deberá tener acceso a una computadora personal, sea en un laboratorio de computadoras o en su casa. Salón de clases equipado con computadoras, pizarras, acceso al Internet y proyector que se pueda conectar a una computadora personal para desplegar visuales en una pantalla electrónica.

X. Estrategias de Evaluación:

| | |
|-------------------------------|-----|
| Exámenes | 50% |
| Asignaciones y pruebas cortas | 25% |
| Proyecto | 25% |

De ser necesario, se realizará evaluación diferenciada a estudiantes con necesidades especiales.

XI. Estrategias de Avalúo:

Se utilizarán estrategias de avalúo tal como ejercicios prácticos en clase, supervisión de trabajo en grupo o individual, exámenes, asignaciones para realizar fuera del salón de clases, se valorará la participación en clase. Se asignará un proyecto final de la clase, para evaluar lo aprendido en el curso. Este proyecto contará con una rúbrica, como medio de avalúo.

XII. Acomodo Razonable.

La Universidad de Puerto Rico (UPR) reconoce el derecho que tienen los estudiantes con impedimentos a una educación post secundaria inclusiva, equitativa y comparable. Conforme a su política hacia los estudiantes con impedimentos, fundamentada en la legislación federal y estatal, todo estudiante cualificado con impedimentos, tiene derecho a la igual participación de aquellos servicios, programas y actividades que están disponibles de naturaleza física, mental o sensorial y que por ello se ha afectado, sustancialmente, una o más actividades principales de la vida como lo es su área de estudios post secundarios, tiene derecho a recibir acomodos o modificaciones razonables. De usted requerir acomodo o modificación razonable en este curso, debe notificarlo al profesor sobre el mismo, sin necesidad de divulgar su condición o diagnóstico. De manera simultánea, debe solicitar a la Oficina de Servicios a Estudiantes con Impedimentos (OSEI) de la unidad o Recinto, en forma expedita, su necesidad de modificación o acomodo razonable.

XIII. Integridad académica:

La Universidad de Puerto Rico promueve los más altos estándares de integridad académica y científica. El Artículo 6.2 del Reglamento General de Estudiantes de la UPR (Certificación 13, 2009-2010, de la Junta de Síndicos) establece que “la deshonestidad académica incluye, pero no se limita a: acciones fraudulentas, la obtención de notas o grados académicos valiéndose de falsas o fraudulentas simulaciones, copiar total o parcialmente la labor académica de otra persona, plagiar total o parcialmente el trabajo de otra persona, copiar total o parcialmente las respuestas de otra persona a las preguntas de un examen, haciendo o consiguiendo que otro tome en su nombre cualquier prueba o examen oral o escrito, así como la ayuda o facilitación para que otra persona incurra en la referida conducta”. Cualquiera de estas acciones estará sujeta a sanciones disciplinarias en conformidad con el procedimiento disciplinario establecido en el Reglamento General de

Estudiantes de la UPR vigente. **Para velar por la integridad y seguridad de los datos de los usuarios, todo curso híbrido, a distancia y en línea deberá ofrecerse mediante la plataforma institucional de gestión de aprendizaje, la cual utiliza protocolos seguros de conexión y autenticación. El sistema autentica la identidad del usuario utilizando el nombre de usuario y contraseña asignados en su cuenta institucional. El usuario es responsable de mantener segura, proteger y no compartir su contraseña con otras personas.**

Política de Integridad Académica de la Universidad de Puerto Rico, Recinto de Río Piedras: Certificación Núm. 64 Año Académico 2022-2023 del Senado Académico: La Universidad de Puerto Rico promueve los más altos estándares de integridad académica y científica. El Recinto de Río Piedras de la Universidad de Puerto Rico (UPRRP) está comprometido con mantener y promover un ambiente intelectual y ético basado en los principios de integridad y rigor académico, confianza, respeto mutuo y diálogo sereno entre las personas de la comunidad universitaria esenciales para el logro de su misión. La integridad implica la firme adherencia a un conjunto de valores éticos fundamentales, tales como la honestidad, el respeto y la responsabilidad. La integridad académica es parte, no solo de la enseñanza y el aprendizaje, sino de las relaciones e interacciones consustanciales al proceso educativo, investigativo y administrativo. Debe permear todos los ámbitos de la vida y la comunidad universitaria. Esta Política de Integridad Académica (de ahora en adelante Política) se sostiene en el quehacer académico compartido entre los integrantes de la comunidad universitaria al promulgar y afianzar estos valores mediante la educación, el diálogo y la prevención. Se enfoca, principalmente, en el ámbito estudiantil en el proceso de enseñanza y aprendizaje y la investigación. Sin embargo, la integridad académica atañe a todos los integrantes de la comunidad universitaria: estudiantes, personal docente y no docente. <https://senado.uprrp.edu/wp-content/uploads/2023/01/CSA-64-2022-2023.pdf>

XIV. ***Política y procedimiento para el manejo de situaciones de discrimen por sexo o género en la Universidad de Puerto Rico:***

La Universidad de Puerto Rico (Universidad), como institución educativa y centro laboral, tiene como norte proteger los derechos y ofrecer un ambiente seguro a todas las personas que interactúan con la institución, ya sean estudiantes, empleados, contratistas y/o visitantes (en adelante “personas concernidas”). En atención a ello, se promulga la presente Política por virtud de la Certificación 107, JG 2021-2022, con el fin de promover un ambiente de respeto a la diversidad y los derechos de los integrantes de la comunidad universitaria. Se establece el protocolo, para el manejo de situaciones relacionadas con las siguientes conductas prohibidas: discrimen por razón de sexo, género, o embarazo, hostigamiento sexual, violencia sexual, violencia doméstica, violencia en cita y acecho, (en adelante, «las conductas prohibidas»), en el ambiente de trabajo y estudio.

XV. **Sistema de Calificación.**

| | |
|----------|---|
| 90 – 100 | A |
| 80 – 89 | B |
| 65 – 79 | C |
| 60 – 64 | D |
| 0 – 59 | F |

XVI. Plan de contingencia en caso de una emergencia.

En caso de surgir una emergencia o interrupción de clases, el profesor continuará ofreciendo el curso utilizando la modalidad a distancia o en línea, según establecidas en este prontuario oficial. De acuerdo a la información oficial y las directrices institucionales, el profesor realizará esfuerzos para comunicarse con los estudiantes vía correo electrónico institucional u otros medios alternos disponibles para coordinar la continuidad del ofrecimiento.

If an emergency or an interruption of courses occurs, course offerings will take place with the support of distance learning modalities, as established in the official syllabus. In compliance with official communications and institutional guidelines, the professor will make efforts to communicate with students via institutional email or other available communication outlets to coordinate the continuity of course work.

XVII. Bibliografía.

Libro de Texto:

Galit Shmueli, Peter C. Bruce, and Nitin R. Patel. 2017. Data Mining for Business Analytics: Concepts, Techniques, and Applications in R. Wiley Publishing
ISBN-13: 978-1118879368

Referencias

- Charu C. Aggarwal (2015) Data Mining: The Textbook. Springer.
- Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal (2016). Data Mining: Practical Machine Learning Tools and Techniques. 4th Edition. Morgan Kaufmann.
- Miroslav Kubat (2017). An Introduction to Machine Learning. Second edition. Springer.
- James D. Miller (2017). Statistics for Data Science: Leverage the power of statistics for Data Analysis, Classification, Regression, Machine Learning, and Neural Networks. Packt Publishing.
- John D. Kelleher, Brian Mac Namee, Aoife D'Arcy. 2015. Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies 1st Edition. MIT Press.

- Brett Lantz (2015) Machine Learning with R: Expert techniques for predictive modeling to solve all your data analysis problems 2nd Edition. Packt Publishing.
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013) An Introduction to Statistical Learning: with Applications in R. 1st ed. 2013, Springer.
- Hadley Wickham , Garrett Grolemund (2017) R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. 1st Edition. O'Reilly Media.
- (2015) Storytelling with Data: A Data Visualization Guide for Business Professionals. Wiley.

Otros Recursos

- The R Project for Statistical Computing, <http://www.r-project.org/>
- “The Knowledge Discovery Mine”, enlaces útiles a otros recursos, publicaciones relacionadas a minería de datos, enlaces a repositorios de datos y un catálogo de software para minería de datos. <http://www.kdnuggets.com/index.html>
- Repositorio de conjuntos de datos <https://www.kaggle.com/datasets>
- Census Bureau Homepage:, <http://www.census.gov>
- U.S. Government’s open data <http://www.data.gov/>
- Puerto Rico open data <https://data.pr.gov/>
- UC Irvine Machine Learning Repository <http://archive.ics.uci.edu/ml/>
- Amazon public Data Sets <https://aws.amazon.com/public-data-sets/>