**Universidad de Puerto Rico**
**Recínto de Río Piedras**
**Facultad de Administración de Empresas[1]**
**Instituto de Estadística[2] 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. Prerequisito: ESTA 3042 Estadística para la Administración de Empresas II, MECU 3032 Métodos Cuantitativos para Administración de Empresas II o equivalentes. 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.

[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.

4) Aplicar técnicas para crear modelos predictivos por medio de técnicas de clasificación supervisada.
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 clasificacion 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

| ***Temas*** | ***Distribución de tiempo (horas)*** |
|---|---|
| I. Introducción<br>a. ¿Qué es minería de datos?<br>b. Aplicaciones de minería de datos.<br>c. Pasos en el proceso de minería de datos.<br>d. Consideraciones éticas en minería de datos. | 1.5 |
| II. Pre-procesamiento de datos<br>a. Tratamiento de datos faltantes<br>b. Reducción de la dimensionalidad:<br>i. Selección de variables<br>ii. Extracción de variables<br>c. Discretización<br>d. Normalización | 9 |
| III. Visualización<br>a. Graficas de barras, scatterplots, boxplots e histogramas.<br>b. “Heatmaps”: Visualizar correlación y valores perdidos<br>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 | 6 |

| | |
|---|---|
| b. Otras gráficas especializadas: Visualización de datos en redes, datos jerárquicos y datos geográficos | |
| IV. Clasificación Supervisada<br>a. Clasificación Bayesiana<br>b. Evaluación de un clasificador<br>c. Clasificación usando vecinos cercanos (KNN)<br>d. Arboles de decisión<br>e. Otros métodos de clasificación: Neural Networks, Support Vector Machines | 9 |
| VI. Reglas de asociación<br>a. Análisis de la canasta de Mercado<br>b. El algoritmo a priori<br>c. De conjuntos de elementos (“itemsets”) frecuentes a reglas de asociación<br>d. Minería de reglas de asociación multidimensional<br>e. De reglas de asociación a análisis de correlación | 6 |
| V. Analítica de Datos<br>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.<br>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, ejerciciós de aplicacion en clase, asignaciónes y discusión de ejercicios. Uso del programa R a traves de todo el curso y otros programas para aplicaciones para tenas 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
Según la Ley de Servicios Educativos para Personas con Impedimentos (Ley 51 del 7 de junio de 1996), todo estudiante que requiera acomodo razonable deberá notificarlo al profesor el primer día de clases. Los estudiantes que requieren acomodo razonable o reciban servicios de Rehabilitación Vocacional deben comunicarse con el profesor al inicio del semestre para planificar el acomodo razonable y el equipo asistido necesario conforme a las recomendaciones de la Oficina de Asuntos para las Personas con Impedimento (OAPI) del Decanato de Estudiantes.

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 Núm. 13, 2009-2010. De la Junta de Síndicos) establece que "la deshonestidad acdémica incluye, pero no se limita a: acciones fraudulentas, la obtención de notas grados académicos valiéndose de falsas o fraudulentas simulaciones, copiar total o parcialmente la labor acdémica de otras personas, 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.

XIV. Normativa sobre discrimen por sexo y género en modalidad de violencia sexual
La Universidad de Puerto Rico prohíbe el discrimen por razón de sexo y género en todas sus modalidades, incluyendo el hostigamiento sexual. Según la Política Institucional contra el Hostigamiento Sexual en la Universidad de Puerto Rico, Certificación Num. 130, 2014-2015 de la Junta de Gobierno, si un estudiante está siendo o fue afectado por conductas relacionadas a hostigamiento sexual, puede acudir ante la Oficina de la Procuraduría

Estudiantil, el Decanato de Estudiantes o la Coordinadora de Cumplimiento con Título IX para orientación y/o presentar una queja.

XV. Sistema de Calificación.

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

XVI. 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/`