Codigo de asignatura: IO253

Nombre del programa académico	Maestria en Investigacion Operativa y Estadística
Nombre completo de la asignatura	**Minería de Datos**
Área académica o categoría	Investigacion de operaciones y estadistica
Semestre y año de actualización	Primer semestre de 2018
Semestre y año en que se imparte	IV-to semestre
Tipo de asignatura	[ ] Obligatoria [x] Electiva
Número de créditos ECTS	6 ECTS
Director o contacto del programa	Jose A. Soto Mejia
Coordinador o contacto de la asignatura	Ignacio Pérez Vélez

Descripción y contenidos

1. Breve descripción

Se presenta una metodología (CRISP—DM) para desarrollar un proyecto de analítica de datos. En el contexto de dicha metodología se desarrollan métodos de preparación de datos, se estudian los aspectos analíticos y conceptuales de los principales modelos en minería orientada y no orientada, se explican los métodos de evaluación de modelos y estrategias organizacionales de despliegue de un proyecto.

2. Objetivos del Programa Académico MIOE (desde la perspectiva de la universidad)

OP2. Presentar las formas de optimizar el uso de los recursos que la empresa utiliza para hacerla más competitiva, aplicando modelos y herramientas de la investigación de operaciones y estadística.

OP3. Presentar técnicas estadísticas cualitativas y cuantitativas multivariadas encaminadas a soportar la toma de decisiones en los campos de la ingeniería teniendo en cuenta el contexto global de la sociedad. OP4.Fomentar la investigación en temas relacionados con las técnicas de investigación de operaciones y la estadística, teniendo en cuenta el rigor ético, moral y científico.

Objetivos de la Asignatura (desde la perspectiva del profesor)

- Presentar una metodologia para resolver un problema de ingenieria mediante analítica de datos.
- Ilustrar los diferentes tipos de problema en minería de datos (orientada y no orientadada).
- Relacionar problemas de ingenieria y problemas de minería de datos.
- Desarrollar los aspectos conceptuales de distintos modelos en analítica.
- Presentar la forma de evaluar distintos modelos para resolver problemas de ingenieria mediante analítica de datos de la forma más adecuada.
- Describir la manera de combinar y utilizar modelos de analítica de datos para resolver problemas de ingenieria.

3. Resultados de aprendizaje (desde la perspectiva del estudiante)

RA1.Identificar cada una de las etapas en la metodología CRISP—DM
RA2.Conocer las tareas a desarrollar y los reportes a producir en cada una de las etapas de CRISP-DM
RA3.Diferenciar problemas de minería orientada y no orientada.
RA4.Identificar el problema de minería que corresponde a una pregunta de negocio.
RA5.Aplicar técnicas estadísticas y de investigación de operaciones en modelos de minería de datos.
RA6.Analizar los indicadores de calidad de un modelo.
RA7.Combinar diferentes modelos de minería de datos para mejorar modelos básicos.
RA8.Aplicar técnicas estadísticas en la preparación de bases de datos analíticas.
RA9.Programar flujos de análisis de datos en herramientas de software libre como KNIME y R.
RA10.Fomentar el trabajo en equipo

4. Contenido

T1: Conceptos básicos de minería de datos. (14 h)
T2: Metodología CRISP-DM (15 h)
T3: Preparación de la base analítica. (15 h)

T4: Métodos básicos de clasificacion :1R, Bayes Naive, K vecinos cercanos, regresión logística. (15 h). T5: El problema de clasificación. Otros métodos: Árboles de decisión, redes neuronales, SVM. (30 h). T6: Construcción de conglomerados: K medias, DBSCAN, metodos difusos y jerárquicos (15 h). T7: Análisis de canasta de mercado . (15 h). T8: Minería de texto. (15 h)
T9: Otras técnicas: Métodos de conjunto. Ada boost, Bosques aleatorios. Deep learning (10 h)

5. Requisitos: Conocimientos en: Algebra Lineal, Teoría de la Probabilidad, Modelos estadísticos lineales, Programación Lineal y no lineal.

6. Recursos

Material guía: Módulos, diapositivas y material suministrado por el profesor.

Textos guía

- James Gareth, Witten Daniela, Hastie Trevor, Tibshirani Robert; "*An Introduction to Statistical Learning with Applications in R*"; Springer 8a impression 2017. (gratuito para estudio personal disponible en https://statlearning.com/)
- P. Kroese, Zdravko I. Botev, Thomas Taimre, Radislav Vaisman; "*Data Science and Machine Learning - Mathematical and Statistical Methods".* CRC, 2020 (gratuito para estudio personal disponible en https://acems.org.au/data-science-machine-learning-book-available-download)

Textos complementarios

- Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola "*Dive into Deep Learning Release 0.15*.1". 2020 (gratuito para estudio personal disponible en https://statlearning.com/) https://d2l.ai/)
- Shai Shalev-Shwartz & Shai Ben-David; "*Understanding Machine Learning - From Theory to Algorithms*". Cambridge University Press, 2014. (gratuito para estudio personal disponible en https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/index.html)
- Molnar, Christoph; "*Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*."; 2021. Solo disponible en foramto electrónico en el git hub del autor https://christophm.github.io/interpretable-ml-book/

Lecturas adicionales

- Lindsay I Smith, A tutorial on Principal Components Analysis, February 26, 2002 (http://www.iro.umontreal.ca/~pift6080/H09/documents/papers/pca_tutorial.pdf)
- Michael Steinbach, Levent Ertöz, and Vipin Kumar, The Challenges of Clustering High Dimensional Data.
- Wickham, Hadley "Tidy Data". Journal of Statistical Software, 20 February 2013.
- Ribeiro Marco Tulio, Singh Sameer, Guestrin Carlos; "*Why Should I Trust You?": Explaining the Predictions of Any Classifier*". 2016 https://arxiv.org/abs/1602.04938
- Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola "*Dive into Deep Learning Release 0.15*.1". 2020 (gratuito para estudio personal disponible en https://statlearning.com/) https://d2l.ai/)
- Shai Shalev-Shwartz & Shai Ben-David; "*Understanding Machine Learning - From Theory to Algorithms*". Cambridge University Press, 2014. (gratuito para estudio personal disponible en https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/index.html)

Software

- KNIME. Libre distribución, disponible en www.knime.org. Complementos de minería de texto, Weka y BIRT. Este es un software de aplicación de uso extendido en la comunidad de Ciencia de Datos. Es un programa que implementa un paradigma gráfico de programación que hace que el estudiante pueda rápidamente estar construyendo modelos para problemas reales. Tiene posibilidades de integrarse con Python, R, Keras y TensorFlow. Cuenta con extensiones para anlítica de texto, imágenes y redes sociales. A lo largo del curso los estudiantes van construyendo los programas (workflows en la terminología del programa) correspondientes a los modelos discutidos en clase. Se trabaja con datos reales disponibles en distintas fuentes como por ejemplo

con datos abiertos en https://herramientas.datos.gov.co/es, o
https://datosabiertos.bogota.gov.co/about
https://www.colombiacompra.gov.co/transparencia/gestion-documental/datos-abiertos

Bases de datos suscritas por la Universidad

7. Herramientas técnicas de soporte para la enseñanza

Clase Magistral en la cual se presenta los temas a desarrollar soportados con diapositivas y, para algunos puntos con lecturas previas de los estudiantes.
Lectura fuera del aula sobre algunos temas seleccionados cono tidy data, big data clustering, árboles de decision y máquinas de soporte vectorial.
Talleres en clase utilizando KNIME con datos reales tomados de repositorios públicos como el de Universidad de California en Irvine (UCI).
Talleres fuera del aula en grupos de 3 estudiantes resolviendo problemas de analítica com datos reales.
Video sobre visión por computador y lenguaje natural (Prof. FeiFei Li en TED).
Vídeo sobre la ética en minería de datos («Weapons of Math Destruction» en TED).

8. Trabajos en laboratorio y proyectos

- Talleres correspondiente a temas T3 a T9; 10 horas (incluidas en las sesiones de clase)
- Talleres prácticos en grupos de 3 estudiantes sobre los temas T4 y T6 ; 6 horas.
- Proyecto final en grupos de 3 estudiantes com una dedicación de 12 horas por estudiante.

9. Métodos de aprendizaje

Trabajo en grupos. Exposiciones magistrales por parte del docente.
Talleres (en el aula y fuera de ella). Solución de problemas de tipo cálculo numérico.

10. Métodos de evaluación

- Examen escrito sobre preparación de datos correspondiente a T1 y T2 (5%) (RA1, RA2, RA3, RA4)
- Evaluación del taller al final del tema T4 (15%): (RA2, RA3, RA4, RA5, RA6, RA8 RA9, RA10)
- Evaluación del taller al final del tema T6 (15%): (RA2, RA3, RA4, RA5, RA6, RA8 RA9, RA10)
- Evaluación de proyecto final (30%): (RA1, RA2, RA3, RA4, RA5, RA6, RA7, RAW, RA9 y RA10)
- Examen sobre los fundamentos conceptuales de los métodos al final del curso. (15%) :(RA3, RA5, RA7). Nota de talleres: T3, T4, T5, T6, T7 y T8: (20%): (RA5, RA7, RA9)