FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIA DE DATOS

PROGRAMA DE CURSO

Curso	**CCCCC – Fundamentos de Analítica 1**
Clasificación	Núcleo Programa
Período académico	2020 - 1
Número de Horas	52 / 4 Créditos
Profesor	**Javier Diaz Cely** – jgdiaz@icesi.edu.co

DESCRIPCIÓN

En este curso se introducen los conceptos básicos de la analítica de datos, presentando las características de los modelos de aprendizaje automático (*machine learning*), desde un enfoque teórico (introductorio) y práctico, distinguiendo entre los modelos supervisados (permitiendo la predicción) y no supervisados (encontrando patrones estructurales en los datos), y estudiando las métricas de calidad de los mismos y los protocolos de evaluación que permiten valorarlos y compararlos.

En el caso de los modelos de aprendizaje supervisado, el curso tratará la importancia del problema de sobreaprendizaje (*overfitting*). Éste está íntimamente relacionado con la complejidad de los modelos e impide generalizar los resultados a conjuntos de datos diferentes a los utilizados para su entrenamiento. Además, da pie a la utilización de protocolos de evaluación como el *holdout* y la validación cruzada, que permiten detectarlo y obtener una mejor valoración de la calidad de los modelos. Se estudiarán los modelos K-NN, árboles de decisión, regresión logística, redes neuronales tradicionales, Bayes Ingenuo, entre otros.

En el caso de los modelos de aprendizaje no supervisado, el curso se enfocará en las técnicas de segmentación de datos (*clustering*, algoritmos de K-Means, Clustering jerárquico), reducción de dimensionalidad (PCA).

OBJETIVO GENERAL

Al finalizar el curso, los estudiantes podrán de aplicar técnicas de clasificación, regresión, clustering y reducción de dimensionalidad para descubrir información valiosa y guiar la toma de decisiones del negocio, descubrir oportunidades estratégicas y mejorar el desempeño organizacional.

OBJETIVOS TERMINALES

- **Objetivo terminal 1.** Formular preguntas relevantes para el negocio, que puedan ser respondidas a partir del análisis de los datos disponibles.
- **Objetivo terminal 2.** Aplicar la metodología ASUM-DM al desarrollo de proyectos de analítica de datos.
- **Objetivo terminal 3**. Reconocer problemas de clasificación, regresión, clustering y reducción de dimensionalidad.
- **Objetivo terminal 4**. Explicar las implicaciones del sobre aprendizaje y los protocolos de evaluación que permiten identificarlo y combatirlo.
- **Objetivo terminal 5**. Aplicar R para el entrenamiento y la evaluación de modelos de analítica, siguiendo las métricas pertinentes al tipo de problema que solucionan.

PROGRAMA DEL CURSO

Unidad 1. Introducción a la analítica de datos

- **Objetivo Específico 1.** Entender las generalidades de la analítica de datos, el *big data* y sus aplicaciones como solución a problemáticas reales
- **Objetivo Específico 2.** Discutir las etapas de la metodología ASUM-DM y cómo se puede aplicar a proyectos de analítica de datos

Unidad 2. Modelos de aprendizaje supervisado

- **Objetivo Específico 1.** Identificar problemáticas que se puedan resolver a partir de modelos de aprendizaje supervisado, ya sea de clasificación o de regresión.
- **Objetivo Específico 2.** Aplicar modelos de K-NN, Bayes ingenuo, regresión logística, árboles de decisión, y ensambles de modelos a conjuntos de datos para responder a preguntas de negocio involucrando modelos de aprendizaje supervisado, utilizando el lenguaje R.
- **Objetivo Específico 3.** Explicar los problemas de sub-aprendizaje o sobre-aprendizaje en modelos supervisados
- **Objetivo Específico 4.** Comparar modelos de aprendizaje supervisado con respecto a métricas de ajuste, utilizando diferentes protocolos de evaluación
- **Objetivo Específico 5.** Establecer los mejores valores de los parámetros de los modelos de aprendizaje supervisado

Unidad 3. Modelos de aprendizaje no supervisado

- **Objetivo Específico 1.** Identificar problemáticas que se puedan resolver a partir de modelos de aprendizaje no supervisado.
- **Objetivo Específico 2.** Aplicar modelos de segmentación de datos (*clustering*) como K-Means, Clustering jerárquico.
- **Objetivo Específico 3.** Aplicar modelos de reducción de dimensionalidad de los datos como PCA.
- **Objetivo Específico 4.** Establecer los mejores valores de los parámetros de los modelos de aprendizaje no supervisado.

METODOLOGÍA

El curso se desarrollará por unidades, de acuerdo al contenido presentado, con espacios de discusión, aplicación y análisis de los conceptos, y la participación activa de los estudiantes.

Los estudiantes deberán preparar, antes de la clase, los temas que asigne el profesor (Guías de lectura), que serán discutidos al comienzo de la clase para aclarar todo tipo de dudas, para ser luego evaluados a partir de quices. Bajo el esquema de trabajo de este curso, preparar un tema significa hacer una lectura crítica (análisis y síntesis) del tema/material de lectura que corresponda, indagar sobre los aspectos desconocidos, resolver las preguntas y los ejercicios planteados, y llegar a clase dispuesto a discutir el tema y a resolver las dudas que hayan surgido al realizar las actividades mencionadas y las propuestas por el profesor. La parte final de cada clase incluye una presentación del tema por parte del profesor para afianzar los conceptos, y una aplicación de los mismos a conjuntos de datos en forma de talleres prácticos.

EVALUACIÓN

Forma de Evaluación	Porcentaje	Sesión en que se realiza (opcional)
Lecturas / Quices / Tareas	20%	
Examen Parcial 1	25%	
Examen Parcial 2	25%	
Trabajo de aplicación	30%	
Total	**100%**	

Los exámenes del curso comprenden los temas en sus aspectos tanto teóricos (en forma de cuestionario de escogencia múltiple, incluyendo además pequeños ejercicios) como prácticos (caso de aplicación).
El trabajo de aplicación se hará sobre un dataset escogido por el profesor, que deberá ser validado previamente por el profesor.

RESUMEN DE HOJA DE VIDA DEL PROFESOR

Director de la Maestría en Ciencia de Datos y Coordinador del Diplomado de Analítica y Big Data en la Universidad ICESI. PhD en Informática de la Universidad de Paris VI, Pierre y Marie Curie (Sorbonne Université, Francia), con títulos de Maestría en Inteligencia Artificial y Reconocimiento de Patrones de la misma Universidad y Maestría en Finanzas Corporativas del Conservatorio Nacional de Artes y Oficios (CNAM, Francia).
Investigador asociado al Centro de Excelencia y apropiación en Big Data y Data Analytics – Alianza CAOBA. Ha trabajado en el sector bancario (Société Générale, Banco Falabella), de telecomunicaciones (Carvajal, France Télécom - Orange) y de consultoría (Altran) en Francia y Colombia (Bancolombia, Nutresa, DNP, SDH de Bogotá, TQ, CELSIA-EPSA, Nueva EPS). Profesor de la Facultad de Ingeniería de la Universidad ICESI, investigador en ciencia de datos del grupo de investigación I2T.
Sus intereses incluyen la aplicación de las técnicas de Machine Learning, Deep Learning y Big Data a las problemáticas encontradas en los sectores financieros, de marketing y de salud. Ha escrito más de una docena de artículos expuestos y publicados en conferencias y journals internacionales.

BIBLIOGRAFÍA

R

1. An Introduction to Statistical Learning with Applications in R. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. Springer. 2009. Gratis on-line (http://www-bcf.usc.edu/~gareth/ISL/)
2. R for Data Science. Hadley Wickham, G. Grolemund,
3. R in Action: Data Analysis and Graphics with R. Robert Kabacoff. Manning, 2014
4. Practical Data Science with R. Nina Zumel, John Mount. Manning, 2014

GENERAL

5. Data Science for Business What you need to know about data mining and data-analytic thinking. Foster Provost, Tom Fawcett. O'Reilly, 2013

Software de referencia: Todas las aplicaciones utilizadas durante el curso son de acceso libre.

- R http://www.r-project.org/
- R Studio http://www.rstudio.com/
- Weka https://www.cs.waikato.ac.nz/ml/weka/downloading.html