Magíster en Gestión y Dirección de Empresas – MBA

IN7121 Data Analytics Aplicado a la Gestión

PROGRAMA DE CURSO

Requisito	**:**	**Autorización**
Créditos	**:**	**03**
Equipo Docente		
Profesor	**:**	**Charles Thraves**
Profesor Auxiliar	**:**	
Semestre	**:**	**Primavera 2023**

1. DESCRIPCIÓN

El curso introduce los principales métodos de aprendizaje automático, i.e., machine learning, en diferentes aplicaciones de la industria. Las herramientas analíticas que se verán en el curso permiten tomar mejores decisiones frente a diversos problemas que enfrentan las organizaciones, generando ventajas competitivas en diversos aspectos tales como ahorro de costo, mayor utilización de recursos, mejores estimaciones de demanda, etc. Cada método se motivará con una aplicación real, indicando las ventajas y desventajas del método bajo estudio respecto a otras técnicas. Se explicará el funcionamiento de cada método, indicando la lógica, sin entrar en detalle sobre la matemática que hay detrás. Se pondrá en práctica cada una de las herramientas con diversos ejemplos con datos reales usando el lenguaje de programación R y Rstudio. Se contempla para este curso siete clases de cátedra más una sesión para el examen. Hay dos tareas, y el proyecto del curso donde los alumnos puedan poner en práctica los conceptos de la clase. De esta forma, una vez finalizado el curso, se espera que el alumno sea capaz de entender las técnicas como también aplicarlas con datos reales. No es requisito para el curso poseer conocimiento en lenguajes de programación.

2. OBJETIVOS

Generales

Se espera que una vez terminado el curso, los alumnos comprendan diferentes herramientas de machine learning entendiendo su lógica como también cuándo son adecuados de aplicar.

Específicos

- Conocer conceptos generales de machine learning aplicables a todos los métodos supervisados, tales como set de entrenamiento, testeo, sobre- o sub-ajuste, validación cruzada, hiper-parámetros.
- Entender la lógica de cada método predictivo como también de cada método no supervisado.
- Aprender a aplicar las técnicas con datos reales en un lenguaje de programación.

3. METODOLOGÍA

El curso consta de 7 clases de cátedra de 3 horas cada más una sesión de 3 horas para el examen. En las clases de catedra se expondrá la materia del curso, la cual se complementará con: reuniones con el auxiliar del curso, tareas y un proyecto. Se espera que los alumnos asistan y participen en clase, la asistencia es obligatoria, (ver más detalles en la sección EVALUACIONES). Las tareas y el proyecto son grupales. Además, cada grupo deberá reunirse en **dos instancias con el auxiliar del curso** para recibir apoyo y guía acerca del proyecto del curso. **Está estrictamente prohibido responder preguntas con personas de otros grupos, o hacer uso compartido de material entregable, y usar material resuelto de internet o versiones anteriores del curso (esto último aplica para tareas, informes). Además, para el proyecto, está prohibido la utilización de proyectos ya resueltos por terceros.**

Bibliografía

- James, G. y D. Witten y T. Hastie y R. Tibshirani (2017) “An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)”, Springer. https://hastie.su.domains/ISLR2/ISLRv2_website.pdf
- Bertsimas, D. y A. O’Hair y ,W. Pulleyblank (2016) “The Analytics Edge”

Bibliografía (opcional)

- Alpayden, E. “Introduction to Machine Learning” Third Edition. PHI https://dl.matlabyar.com/siavash/ML/Book/Ethem%20Alpaydin-Introduction%20to%20Machine%20Learning-The%20MIT%20Press%20(2014).pdf
- Christopher M. Bishop, “Pattern Recognition and Machine Learning (Information, Science, and Statistics)”, Springer https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf
- Hastie, T. y R. Tibshirani y J. Friedman (2009), “The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)” Springer. https://hastie.su.domains/Papers/ESLII.pdf

4. EVALUACIÓNES

Las evaluaciones corresponden a tareas, informes, participación, y asistencia. La ponderación de cada evaluación en la nota final del curso se indica entre paréntesis.

- Tareas (40%):
- Participación (10%)
- Proyecto (50%): Cada grupo deberá llevar a cabo un proyecto donde se apliquen las herramientas vistas en el curso. Durante el semestre se deberán entregar dos informes de avance del proyecto. En la última clase del curso, se realizará una presentación del proyecto más una entrega del informe final.

Tareas: Habrán dos tareas en donde se deberán aplicar los contenidos del curso en un contexto aplicado. Las tareas son en grupos de a 3 personas.

- Tarea 1: Fecha de entrega **Miércoles 27 de Diciembre**
- Tarea 2: Fecha de entrega **Miércoles 17 de Enero**

Participación: Se considerará la participación en clases de los alumnos. En particular: ¿Está escuchando atentamente a la clase? ¿Son las acotaciones concisas? ¿Son las intervenciones relevantes a la discusión en clase? ¿Se muestra en los comentarios un análisis de lo que se está discutiendo? ¿Hay intención de participar? ¿Tiene la cámara encendida en caso de estar online?

Proyecto: Los proyectos son en grupos de a 3. Cada grupo deberá llevar a cabo un proyecto donde se apliquen las herramientas vistas en el curso. Durante el semestre se deberán entregar dos informes de avance del proyecto. Al final del curso, se realizará una presentación del proyecto más una entrega del informe final. La presentación se realizará en la última clase, en donde todos los miembros del grupo deben presentar el problema abordado, en particular mencionar el contexto, los datos usados, la variable a predecir (en caso de que exista), las covariables usadas, las herramientas predictivas y/o analíticas usadas, resultados, y conclusiones.

Los integrantes de cada grupo deben buscar una aplicación basada en datos reales donde aplicar las herramientas del curso. Pueden ser datos obtenidos de las organizaciones en las cuales están trabajando, o bien pueden ser datos obtenidos de otros contextos en donde tenga sentido aplicar las herramientas que se verán en el curso.

Habrá sesiones de reunión con un auxiliar quien apoyará con el avance de los proyectos.

5. CONTENIDOS Y CRONOGRAMA

Clase 1, Miércoles 6 de Diciembre
Tema: Introducción a Machine Learning, conceptos básicos, Introducción al lenguaje de programación R
Descripción: Introducción a los conceptos básicos de machine learning, distinguiendo entre los tipos de problemas y respuestas a los que se enfrentan los métodos supervisados y no supervisados. Se revisará el concepto de sub- y sobreestimación de modelos predictivos. Se distinguirá entre lo que es el set de entrenamiento, validación, y testeo. Adicionalmente, se revisará el concepto de regularización. En la segunda parte de la sesión de revisarán algunos conceptos básicos de estadística tales como promedio, varianza, desviación estándar, covarianza, y correlación. Finalmente, se introducirá el lenguaje de programación R.

Preparación:
- Instalar el lenguaje de programación R y RStudio. Instrucciones en https://rstudio-education.github.io/hopr/starting.html

Requisito: Sin requisito

Evaluación: No hay

Bibliografía:

Modalidad: Híbrida

Clase 2, Miércoles 13 de Diciembre
Tema: Regresión Lineal
Descripción: Se verá el método de regresión lineal con una aplicación en la industria del vino y la salud. Se aprenderá a leer el output de una regresión lineal: coeficientes, p-valores, residuos, coeficiente de determinación, multicolinearidad. Además, se revisarán métodos alternativos a la regresión lineal tales como la regresión LASSO, introduciendo el concepto de hiper-parámetro, y su conexión con los sets de entrenamiento y validación.

Preparación:
- Ver video https://www.youtube.com/watch?v=8WMRj9mTQtI
- Leer https://www.forbes.com/sites/randybean/2022/09/18/moneyball-20-years-later-a-progress-report-on-data-and-analytics-in-professional-sports/?sh=5da2c86773d9

Requisito:
- Entrega primer informe de avance del Proyecto

Evaluación: No hay

Bibliografía:
- Bertsimas, D. y A. O'Hair, W. Pulleyblank (2016), *The Analytics Edge*, Capítulo: 1.1, 4.1
- James, G. y D. Witten y T. Hastie y R. Tibshirani (2017), *An Introduction to Statistical Learning*, Capítulo: 1

Modalidad: Híbrida

Clase 3, Miércoles 20 de Diciembre
Tema: Regresión Logística (Logit)

Descripción: Se introducirá el concepto de regresión logística, identificando las condiciones necesarias y contextos bajo los cuales esta herramienta predictiva es propicia a ser utilizada. Se revisará la aplicación práctica del modelo logit en datos reales en R, comprendiendo el output (coeficientes, p-valores), cómo también el concepto de matriz de clasificación, precisión (*precision*), especificidad (*recall)*, y curva ROC*.* Se revisará el modelo de regresión logística multinomial para predicciones de más de dos clases.

Requisito: Sin requisito

Evaluación: no hay

Bibliografía:
- Bertsimas, D. y A. O'Hair, W. Pulleyblank (2016), *The Analytics Edge*, Capítulo: 7
- James, G. y D. Witten y T. Hastie y R. Tibshirani (2017), *An Introduction to Statistical Learning*, Capítulos: 4.1, 4.2, 4.3 (sin 4.3.2)

Modalidad: Híbrida

Clase 4, Miércoles 27 de Diciembre

Tema: Árboles de Clasificación y Regresión, y Random Forest

Descripción: Se introducirá el concepto de árboles de clasificación y regresión. Se explicará su lógica y cómo esta se diferencia de las técnicas vistas en las clases anteriores. En particular, se revisarán las ventajas y desventaja como también sus limitaciones. Se conocerán los hiper-parámetros más conocidos en este tipo de modelo predictivo. Se aplicará el método en datos reales, y se repasarán conceptos relativos al ajuste del modelo*.*

Requisito: Sin requisito

Evaluación: Entrega de Tarea 1

Bibliografía:
- James, G. y D. Witten y T. Hastie y R. Tibshirani (2017), *An Introduction to Statistical Learning*, Capítulo: 8.1 y 8.2

Modalidad: Híbrida

Clase 5, Miércoles 3 de Enero

Tema: K-vecinos más Cercanos (KNN) – Sistemas de Recomendación

Descripción: Se verá el método KNN, explicando la lógica y funcionamiento tanto para problemas de regresión y clasificación de dos o más clases. Luego se aplicará el método para construir un sistema de recomendación basado en el rating de diferentes películas.

Requisito:

- Llenar la encuesta de rating de películas. Para cada pregunta, responder qué tanta afinidad tienes con la película en una escala de 1 a 5 donde: 1=Me disgustó totalmente, 2=Me disgustó, 1=Me fue indiferente, 4=Me gustó, y 5=Me gustó totalmente. Si no has visto la película. Link de la encuesta: https://forms.gle/vHC4zpkUk7ZzH1J6A
- Leer https://www.forbes.com/sites/insights-teradata/2019/10/01/how-pandora-knows-what-you-want-to-hear-next/?sh=74b762883902

Discutiremos en clases:

- En retrospectiva, ¿cómo evoluciona el mercado de *streaming* durante la década de 2010 a 2020? En su opinión, ¿cómo cree que le irá a este mercado en los próximos años?
- ¿Cuál es la compensación, es decir, los beneficios y los costos, al aumentar el grupo de canciones analizadas? ¿Cómo reaccionaron las bandas pequeñas ante la aparición de estos modelos de negocio?
- ¿Cómo puede una estación de radio por Internet transmitir canciones de manera personalizada a su audiencia? ¿Qué harías para recomendar una canción a un usuario? En otras palabras, qué elementos tomarías en consideración y qué harías con estos.
- ¿Qué otras aplicaciones también podrían beneficiarse de un sistema de recomendación? Para una de estas industrias/aplicaciones, ¿qué atributos usaría para definir alguna métrica de "cercanía" entre dos elementos diferentes?

Trabaje con su grupo en estas preguntas y proporcione un resumen ejecutivo con su respuesta. Esté preparado para discutir las preguntas durante la clase.

- Instalar R y R-Studio
- Instalar librarías: **class, dplyr**
  Para instalar una librería, por ejemplo “class”, correr el comando "install.packages('class')"

Evaluación: no hay

Bibliografía:

- James, G. y D. Witten y T. Hastie y R. Tibshirani (2017) “An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)”, Springer. https://hastie.su.domains/ISLR2/ISLRv2_website.pdf Capítulos: 2.2.3, 3.5, 4.7.6

Modalidad: Híbrida

Clase 6, Miércoles 10 de Enero

Tema: K-medias (*K-means*), Agrupamiento Jerárquico (*Hierarchical Clustering*)

Descripción: Se verán lo métodos de segmentación *K-means* y *hierarchical clustering*, comprendiento las lógicas de cada método. Se darán a conocer diferentes formas de escoger el número de grupos. Para el caso d *hierarchical clustering*, se verán los diferentes resultados al usador distintas métricas de distancia.

Requisito:
- Entrega de segundo informe de avance del proyecto

Evaluación: no hay

Bibliografía:
- James, G. y D. Witten y T. Hastie y R. Tibshirani (2017), *An Introduction to Statistical Learning*, Capítulo: 10.3

Modalidad: Híbrida

Clase 7, Miércoles 17 de Enero

Tema: Análisis de Componentes Principales (ACP)

Descripción: Analizaremos múltiples aplicaciones reales de ACP y cómo se puede usar este método para ayudar a visualizar datos y comprender las principales dimensiones subyacentes de los datos que explican la mayor parte de la variabilidad de éstos. Además, repasaremos algunos conceptos básicos de estadística, mientras que también cubriremos aspectos importantes cuando se trabaja con datos como valores atípicos, transformación y normalización de variables. Se ilustrará cómo aplicar el ACP con datos de una encuesta en diferentes temas políticos, identificando las principales dimensiones que resumen las preguntas de la encuesta.

Preparación:
- Llenar la encuesta en el siguiente link: https://forms.gle/J1Cd59SD2vTzGaPt5
- Instalar R and R-Studio
- Instalar librarías: ggfortify, corrplot, stringr.
  Para instalar una librería, por ejemplo “class”, correr el comando "install.packages('class')"
- Entrega tarea 2

Evaluación: no hay

Bibliografía:
- James, G. y D. Witten y T. Hastie y R. Tibshirani (2017), *An Introduction to Statistical Learning*, Capítulo: 10.2

Modalidad: Híbrida

Clase 8, Miércoles 24 de Enero

Tema: Examen

Descripción: Presentación de grupos

Requisito:
- Entrega del informe final del proyecto

Evaluación:
- Presentación del proyecto y del informe final

Bibliografía:

Modalidad: Híbrida