

## **1. Horario atención a estudiantes, correos electrónicos y nombres de los profesores complementarios**

### **Clase magistral**

Profesor: Andrés Clavijo (andres.clavijo@gmail.com)

Horario de atención a estudiantes: Viernes 4:30 p.m. – 6:00 p.m. o cita previa.

Lugar de atención a estudiantes: W – Piso 7.

Salón: ML – 108b

## **2. Introducción y descripción general del curso**

El curso presentará al estudiante algunas de las técnicas de ciencias de datos más importantes en el análisis. Se comenzará mostrando la explosión de datos que el mundo actual vive a causa de las nuevas tecnologías de la información y las comunicaciones. Nos comenzaremos preguntando ¿cuándo hay un problema de Big Data? ¿Cuáles son los límites y posibilidades de aplicación de Big Data? ¿Cuáles son los efectos de esta revolución en la privacidad y normas sobre protección de datos personales? ¿Qué tecnologías son necesarias para el desarrollo de análisis de grandes datos? Luego estudiaremos los métodos para la analítica de datos, con el fin de comenzar con la exploración y análisis de datos con el paquete estadístico R. Finalizaremos con los métodos avanzados para la analítica de datos.

A lo largo del curso los estudiantes desarrollarán laboratorios en clase que les permita poner en práctica los conceptos aprendidos.

## **3. Objetivos de la materia**

El curso tiene por objetivo brindar a los estudiantes nuevas herramientas y métodos para la analítica de grandes datos. Al final del curso los estudiantes sabrán identificar y diferenciar los conceptos de machine learning, data mining, analítica predictiva y econometría. A su vez, podrán entender los diferentes tipos de estructuras de datos para proponer buenos métodos de análisis y visualización de información. Con respecto al entorno computacional, los estudiantes conocerán del hardware y software más usados para trabajar grandes volúmenes de información.

## 4. Organización del curso

El curso está dividido en cuatro partes.

### **Parte 1. Introducción a la analítica en el contexto del Big Data.**

Introducción al programa.

Laboratorio 1. Exploración de datos.

¿Cuándo hay un problema de Big Data?

- Letouzé (2015)
- Manyika (2011)
- Manske (2016)

Instrucciones para el proyecto final.

### **Parte 2. Entorno informático en la era del Big Data.**

Hardware.

Sistemas distribuidos.

Herramientas analíticas.

- Dean (2014). Capítulos 1, 2 y 3

### **Parte 3. Entorno a los datos.**

Método para la analítica de datos.

- DSBDA (2015). Capítulo 2.

Tipos y calidad de datos.

Introducción al paquete estadístico R.

- Kabacoff (2015). Capítulos 1 y 2.
- James (2015). Pp. 42 – 49.

Exploración y análisis de datos con R.

- Kabacoff (2015). Capítulo 7.

Métodos estadísticos para la evaluación

- DSBDA (2015). pp. 101 – 114.

Laboratorio 2: Estadísticas descriptivas en R.

Técnicas de visualización de información.

- Börner (2015). Pp 44 – 65.

Laboratorio 3: visualizando datos.

### **Parte 4. Método avanzado para la analítica de datos.**

Método de regresión

- Kabacoff (2015). Capítulo 8 y 9.
- DSBDA (2015). Capítulo 6.
- Wooldridge (2003). Capítulo 17.

Laboratorio 4: Métodos de regresión

Análisis espacial

- <https://cran.r-project.org/doc/contrib/intro-spatial-rl.pdf>

Laboratorio 5: Análisis espacial con R/ArcGis

Clusters (Métodos no supervisados de machine learning)

Laboratorio 6: Métodos de cluster

- Kabacoff (2015). Capítulo 16.\*

Clasificación (Métodos supervisados de machine learning)

Laboratorio 7: Métodos de clasificación

- Kabacoff (2015). Capítulo 17.\*
- James (2015). Capítulo 4.

Análisis de textos

Laboratorio 8: Análisis de textos.

## **Presentación de proyectos.**

### **5. Metodología**

El profesor realizará en la primera mitad de la clase presentaciones en power point, que son complementadas por lecturas y vídeos. En la medida que la sesión lo requiera se presentarán vídeos que, por lo general, no tendrán una duración mayor a 5 minutos. En la segunda mitad de la clase, el profesor guiará a los estudiantes en los laboratorios. Los laboratorios son talleres, ejercicios, guías y casos prácticos que permiten a los estudiantes afianzar sus conocimientos. En otras palabras, el método es aprendiendo haciendo.

Los estudiantes deben preparar las clases con las lecturas asignadas para que resuelvan sus dudas en clase.

El objetivo de los laboratorios es aterrizar la parte teórica a la práctica. Es por eso que se recomienda a los estudiantes participar activamente de estos laboratorios

### **6. Competencias**

Al final del curso, los estudiantes estarán en capacidad de:

- a. Manejar la interfaz gráfica de R.
- b. Explorar las distintas herramientas que ofrece R para el análisis de datos.
- c. Crear nuevas visualizaciones para el análisis de datos.
- d. Aplicar distintos métodos de analítica de datos según el tipo de problema y de estructura de datos.
- e. Entender los límites y posibilidades del Big Data.

### **7. Criterios de evaluación (Porcentajes de cada evaluación)**

La evaluación del curso se basará en los siguientes criterios:

Laboratorios (8 laboratorios) (60%)

- Cada laboratorio tiene un peso de (7.5%)

Proyecto final (40%)

El proyecto final consistirá en proponer una solución al reto o problemática planteada por el profesor el primer día de la clase. Los criterios a tener en cuenta son los siguientes:

- Cumplimiento del ejercicio: Las propuestas deben responder a las preguntas planteadas y/o proveer una solución adecuada a la problemática descrita.
- Rigurosidad analítica: los estudiantes deberán presentar el modelo analítico utilizado justificando su uso y explicando por qué es el mejor. Se deberá tener en cuenta los factores que puedan afectar las conclusiones, incluyendo variables que no se tuvieron en cuenta pero que pueden influir en los resultados. Es importante, tener en cuenta el nivel de significancia y la representatividad de los datos utilizados.
- Innovación: Las conclusiones deben ser novedosas y no repetir soluciones ya implementadas.
- Presentación de los resultados: Los resultados deben ser comprensibles incluso para personas que no sean especialistas en el manejo de datos. El proceso de análisis debe ser claramente explicado así como las conclusiones y las implicaciones de éstas.

El proyecto se entregará el último día de clase y se deberá presentar al profesor y a los estudiantes en el salón de clase. El proyecto constará de un documento de no más de 10 páginas y una presentación de no más de 10 slides. Tanto en la presentación como en el documento se deberá describir la temática seleccionada, los datos utilizados, la metodología planteada, los resultados encontrados y la solución al reto o problema, apoyándose de tablas, visualizaciones y demás herramientas vistas en clase.

## **8. Sistema de aproximación de notas definitiva**

Las notas se aproximarán a la centésima con la siguiente excepción:

Nota final entre 2.90 y 2.99 se aproxima a 3.0 si el estudiante ha presentado y aprobado el proyecto final.

**Reclamos** (artículo 62 y 63 del reglamento general de estudiantes de pregrado):

**ART. 62.** Todo estudiante que desee formular un reclamo sobre las calificaciones de cualquier evaluación o sobre la nota definitiva del curso deberá dirigirlo por escrito y debidamente sustentado al profesor responsable de la materia, dentro de los ocho (8) días hábiles siguientes a aquel en que se dan a conocer las calificaciones en cuestión. El profesor dispone de diez (10) días hábiles para resolver el reclamo formulado; vencido el término informará al estudiante la decisión correspondiente.

**ART. 63.** Si el estudiante considera que la decisión no corresponde a los criterios de evaluación, podrá solicitar la designación de un segundo calificador mediante un escrito debidamente sustentado, dirigido al consejo de facultad, dentro de los ocho (8) días hábiles siguientes al conocimiento de la decisión. Si el consejo encuentra fundada la solicitud, procederá a designar, solamente para tal efecto, un segundo calificador cuya decisión debidamente sustentada será definitiva e inmodificable. En ningún caso, el segundo calificador podrá desmejorar la nota inicialmente asignada por el profesor.

### **Fraude**

Según el artículo 109 del reglamento general de estudiantes de pregrado el fraude académico y/o ayuda a cometer fraude son faltas disciplinarias sancionadas con una nota de 0.0 en la prueba en cuestión. En caso de desacuerdo o si se repite la falta se llevará el caso al Consejo de la Facultad de Economía para estudio.

## **9. Bibliografía**

Börner K. 2015. Atlas of Knowledge. Anyone Can Map. MIT Press.

C.B.B.D. Manyika, "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Institute, 2011. <http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation>

Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data. 2015. EMC Education Services. Wiley. <http://www.amazon.com/Data-Science-Big-Analytics-Discovering/dp/111887613X> (sigla DSBDA)

Dean, Jared. 2014. Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. Wiley.

Erl, T., Khattak, W., Buhler, P. 2016. Big Data Fundamentals: Concepts, Drivers & Techniques. Prentice Hall.

James, G., Witten, D., Hastie, T., Tibshirani, R. 2015. Introduction to Statistical Learning. Springer. <http://www-bcf.usc.edu/~garth/ISL/ISLR%20Sixth%20Printing.pdf>

Letouzé, E. 2015. Big Data and Development: General Overview Primer. Data-Pop Alliance White Paper Series. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative. <http://datapopalliance.org/wp-content/uploads/2015/12/Big-Data-Dev-Overview.pdf>

Manske, J., Sangokoya, D., Pestre, G., Letouzé, E. 2016. Oportunidades y requerimientos para aprovechar el uso de Big Data para las estadísticas oficiales y los Objetivos de Desarrollo Sostenible en América Latina". Data-Pop Alliance (Harvard Humanitarian Initiative, MIT Media Lab y Overseas Development Institute). <http://datapopalliance.org/wp-content/uploads/2016/05/Data-Pop-Alliance-LAC-NSO-ES.pdf>

Kabacoff, R. 2015. R in Action. Second Edition. Manning.

Wooldridge, J.M. 2003. Introductory Econometrics. A Modern Approach. Mason, OH, Thomson.

**Último día para subir notas finales en banner: 29 de julio 2016**