PLAN 2009
ANALITICA DE BIG DATA (ELECTIVA)

CARRERA: LICENCIATURA EN CIENCIAS MENCIÓN MATEMÁTICA ESTADÍSTICA

I. IDENTIFICACIÓN

1. Código : 78M
2. Horas Semanales de Clase : 4
   - 2.1. Teóricas : 2
   - 2.2. Prácticas : 2
3. Crédito : 3
4. Pre-Requisito : Ninguno

II. JUSTIFICACIÓN

El Big Data es una forma de aplicar la ciencia de datos, específicamente cuando se requiere explorar grandes cantidades de datos que son generados continuamente y con gran velocidad al interior de las empresas o instituciones.

Estos datos presentan características particulares que impiden su análisis con las técnicas tradicionales, y es así como surgen y evolucionan las nuevas técnicas, entre ellas las conocidas como “Analítica de Big Data”.

Estas técnicas proporcionan una manera de acceder, gestionar y analizar estos datos atendiendo sus particularidades, de modo a encontrar patrones, relaciones y verificar algunas presunciones al respecto de ellas.

III. OBJETIVOS:

1. Identificar algunos de los programas informáticos que permiten encarar el análisis estadístico del Big Data.
2. Identificar los pasos y las herramientas para realizar la recopilación y gestión de datos con las características particulares del Big Data.
3. Emplear las técnicas adecuadas para visualizar y explorar el Big Data
4. Identificar y utilizar algunas herramientas para la técnica de aprendizaje automático (Machine Learning)
5. Describir e Interpretar los resultados obtenidos mediante la utilización de las técnicas para el modelado estadístico en Big Data.

IV. CONTENIDOS

A. UNIDADES PROGRAMÁTICAS

Unidad I: Manejo y gestión de datos en Big Data – Alternativas con R-project y Python.

Unidad II: Técnicas para la visualización y exploración de datos en Big Data

Unidad III: Introducción a las técnicas del Machine Learning en Big Data.

Unidad IV: Descripción, Interpretación y selección de resultados y modelos estadísticos en Big Data

B. DESARROLLO DE LAS UNIDADES PROGRAMÁTICAS

1. Unidad I: **Manejo y gestión de datos en Big Data – Alternativas con R-project y Python.**

1.1. El Programa R-project

1.1.1. El Programa R-project

1.1.2. Instalación e interface

1.1.3. Carga y uso de paquetes

1.1.4. Operaciones y gráficos R

1.1.5. Gestión de archivos y bases de datos con R

1.1.6. Análisis exploratorios de datos con R

1.1.7. Los paquetes de R para Big Data

1.2. El Programa Python

1.1.8. Instalación e interface

1.1.9. Carga y uso de paquetes

1.1.10. Operaciones y gráficos

1.1.11. Gestión de archivos y bases de datos con Python

1.1.12. Análisis exploratorios de datos con Python

1.1.13. Los paquetes de Python para Big Data

2. Unidad II: **Técnicas para la visualización y exploración de datos en Big Data**

2.1. Tipo de variables y distribución

2.2. Distribución de variables de respuesta discreta

2.3. Exploración de datos

2.4. División de los datos en entrenamiento y test

2.5. Preprocesado de los datos

2.6. Imputación de valores ausentes

2.7. Variables con varianza a cero

2.8. Estandarización y escalado

2.9. Binarización de variables cualitativas

2.10. Selección de predictores

3. Unidad III: **Introducción a las técnicas del Machine Learning en Big Data.**

3.1. Creación de modelo predictivo
3.2. K-Nearest Neighbor (KNN)
3.3. Naive Bayes
3.4. Regresión Logística
3.5. Analisis Discriminante Lineal (LDA)
3.6. Support Vertor Machine
3.7. Redes Neuronales (NNET)
3.8. Gradient Boosting
3.9. Arbol de clasificación simple
3.10. Random Forest

4. Unidad IV: **Descripción, Interpretación y selección de resultados y modelos estadísticos en Big Data**

4.1. Interpretación y selección de resultados y los modelos para generar información estadísticas en Big Data

V. METODOLOGÍA

- Exposición
- Demostración
- Planteamiento y solución de situaciones problemáticas
- Trabajo individual y/o grupal
- otros

VI. MEDIOS AUXILIARES

- Equipos informáticos y Software
- Laboratorio de informática
- Material bibliográfico
- Multimedios
- Guía de trabajo

VII. EVALUACIÓN

- La evaluación se regirá conforme al reglamento de la FACEN vigente.

VIII. BIBLIOGRAFÍA BÁSICA

- Ahumada, J. A. (2003). R. para principiantes. University of Hawaii
- Wickham, H. (2016). Ggplot2: elegant graphics for data analysis. Springer.
- Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer

COMPLEMENTARIA

- Dalgaard, P. (2008). Introductory statistics with R. Springer Science & Business Media.
- Chambers, J. (2008). Software for data analysis: programming with R. Springer Science & Business Media.
- http:/topepo.github.io/caret/index.html
- Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.