Diplomado en Herramientas de Minería de Datos para Analítica Empresarial

Versión en Línea

Coordinador académico: M.I. Rafael Gregorio Gamboa Hirales

Nota: Este diplomado es en la modalidad en línea a través de la herramienta Zoom. Se requiere que el participante cuente con computadora, laptop, tablet, teléfono inteligente o cualquier otro dispositivo que permita reproducir audio y video y una buena conexión a internet.

Las clases serán en tiempo real en los días y horario publicados. Las sesiones no serán grabadas y el participante sólo tendrá acceso a las sesiones del diplomado en el grupo al cual se haya inscrito.

Objetivo general

El diplomado proporcionará a los participantes las herramientas, técnicas, habilidades y conocimientos que necesitan para elaborar modelos de analítica enfocados al ámbito empresarial. Se estudiarán los modelos supervisados y los no supervisados, así como las técnicas para la transformación de datos. Con los modelos elaborados se realizarán nuevos procesos a partir de los resultados para perfeccionar el modelado en los términos de los objetivos de negocio. Por ejemplo, en los clasificadores se determinará la calificación de corte para los casos en que haya utilidades y costos asociados a la decisión. Se utilizará Python como herramienta general de procesamiento de datos y formación de modelos, con bibliotecas en este lenguaje o interfaces. Asimismo, se usará Weka por su versatilidad y facilidad para el desarrollo rápido de prototipos y modelos. Se contemplará la puesta en producción en Python y Weka y se estudiará R como alternativa. En la revisión de las herramientas se examinarán los tipos de variables para determinar la idoneidad de la herramienta y las transformaciones requeridas o posibles para optimizar la asertividad del modelo.

¿A quién va dirigido?

Los participantes deben contar con habilidades en el manejo de algún sistema operativo: saber ejecutar programas, manejar archivos y carpetas, comprimir y descomprimir archivos, consultar y bajar archivos de la red, manejar hojas electrónicas y tener nociones de bases de datos relacionales. Se requieren conocimientos básicos de álgebra, álgebra lineal, probabilidad y estadística, y nociones elementales de cálculo diferencial.

Modalidad de enseñanza
Se expondrán los temas y se llevarán a cabo prácticas con tablas de volúmenes considerables de registros. Los trabajos se harán normalmente en computadora. Además, habrá prácticas y tareas para realizar fuera de clase, de modo que los participantes deben instalar en su computadora las herramientas de software. Al final del módulo IV los participantes formarán equipos para emprender un proyecto y presentarlo como trabajo final del diplomado al terminar el módulo V.

Módulo I
HERRAMIENTAS DE PROCESAMIENTO GENERAL DE DATOS

Objetivo
Se proporcionará a los participantes las herramientas básicas de programación para el tratamiento detallado de datos y para manejar arreglos, series y dataframes en Python. Con ello se establecerá una plataforma general para la adquisición y procesamiento de datos. Se analizarán los métodos generales para procesar tablas por medio de dataframes en pandas.

Temario
1. Programación en Python, datos, colecciones y su procesamiento. Iteraciones explícitas e implícitas
2. Numpy, pandas, *seaborn* y matplotlib
3. Estructura del *dataframe*. Índices y columnas
4. Importación y exportación de datos. Codificación
5. Métodos de consolidación de información. Limpieza y transformación de datos
6. Selección de registros. Funciones lambda. Mutación de valores. Concatenación y fusión. Agrupamientos y tablas pivote
7. Visualización de datos
8. Herramientas de raspado web
9. Herramientas para el análisis exploratorio de datos. Ejecución de R y Python en RStudio
10. Generación de documentos con el resultado de las ejecuciones. RMarkdown

Módulo 2
MODELOS SUPERVISADOS

Objetivo
Los participantes comprenderán los fundamentos, hipótesis y finalidad de los métodos supervisados, la manera en que se establecen los objetivos de negocio en términos de la variable objetivo y las métricas que se utilizan para evaluar la bondad de ajuste de los modelos desarrollados. En los clasificadores, se analizará el comportamiento de la calificación de pertenencia del caso a cada una de las clases y su impacto en términos de negocio.

Temario

1. Introducción a la minería de datos. Modelos supervisados y modelos no supervisados
2. Clasificadores. Entropía e información. Información ganada
3. Métricas de bondad de ajuste. Característica operativa del receptor y curva de recuperación de precisión. Concepto de calificación de corte
4. Submuestras de entrenamiento y de prueba. Validación cruzada
5. Bayesiano ingenuo
6. Árboles de clasificación y regresión
7. Regresión logística
8. Máquinas de soporte vectorial
9. Redes neuronales como clasificadores
10. Algoritmos K de vecinos próximos como clasificadores
11. Sobrecarga y compensación
12. Predictores. Objetivo
13. Modelos lineales. Correlación lineal
14. Regresión lineal. Medidas de bondad de ajuste
15. Modelos no lineales. Redes neuronales como predictores
16. Árboles para regresión
17. Algoritmos K de vecinos próximos como predictores
18. Series de tiempo: modelos AR, MA y ARIMA

Módulo 3
MODELOS NO SUPERVISADOS Y TRANSFORMACIÓN DE DATOS

Objetivo

Los participantes asimilarán los fundamentos y objetivos de los modelos de asociación y aprenderán a derivar y aplicar las reglas de asociación. Se analizarán las métricas correspondientes a estos modelos y sus implicaciones para el negocio. Se revisarán las principales técnicas para elaborar modelos de agrupamiento y los criterios para formar el número adecuado de grupos. Se investigará cómo con ayuda de los clasificadores se analiza el contenido de los grupos resultado de un agrupamiento dado y el impacto para el negocio.

Temario

1. Definición, motivación y panorama del aprendizaje no supervisado
2. Modelos de asociación: *a priori,* Eclat, FPGrowth
3. Modelado y descripción de grafos
4. Minería e interpretación de grafos
5. Agrupamiento: dendrogramas y método de Ward
6. Agrupamiento: K de vecinos próximos y DBSCAN. Métricas
7. Reducción de dimensionalidad: análisis de los principales componentes

8. Reducción de dimensionalidad: incrustación de vecinos estocásticos distribuidos en *t*
9. Reducción de dimensionalidad: proyecciones
10. Mapas autoorganizativos
11. Detección de anomalías
12. Redes neuronales no supervisadas (autosupervisadas).

Módulo 4
PROCESAMIENTO Y ANÁLISIS DE TEXTO

Objetivo
Se examinarán los problemas que surgen al elaborar modelos para tratamiento de texto. Se configurarán procesamientos para extraer la información para el análisis de textos y se utilizará esta información con los modelos supervisados y no supervisados vistos en los módulos previos. Los participantes comprenderán la finalidad del análisis de texto en función de sus objetivos.

Temario
1. Captura de datos en páginas web
2. Captura de tuits
3. Corpus. Proceso general de detección de asociación de términos. Medidas de relevancia
4. Matriz de documentos y términos. Búsqueda de narrativas
5. Expresiones regulares para las narrativas
6. Variables indicadoras de narrativas para clasificadores, predictores, asociaciones o conglomerados. Puesta en producción de detección de narrativas en textos
7. Herramientas disponibles para lenguaje hablado. Explotación de datos para *chabots*
8. Bolsa de palabras. Frecuencia de término y frecuencia inversa de documento
9. N-gramas
10. Incrustaciones
11. Análisis de sentimientos

Módulo 5
SISTEMAS DE RECOMENDACIONES. USO DE HERRAMIENTAS EN LA NUBE

Objetivo
Los participantes entenderán la importancia de estudiar los sistemas de recomendaciones, pues constituyen una de las herramientas comerciales más importantes de la actualidad, y aprenderán a elaborar los modelos correspondientes.

Temario

1. Sistemas de recomendaciones
2. Análisis de “clientes” y su comportamiento
3. Conformación de agrupamientos y asociación de “ítems”
4. Procesamiento en flujo
5. Ejemplos en R para proceso de datos
6. Bibliotecas equivalentes para modelos supervisados y no supervisados
7. Procesamiento de matrices dispersas
8. Datos en la nube y herramientas en nube
9. Prácticas con herramientas en nube para elaboración de modelos
10. Avance y presentación del proyecto

Coordinador Académico

M.I. Rafael Gregorio Gamboa Hirales

Rafael Gregorio Gamboa Hirales cursó la licenciatura en Física y Matemáticas en la Escuela Superior de Física y Matemáticas del Instituto Politécnico Nacional y la maestría en Ingeniería en Telecomunicaciones en la Universidad Politécnica de Madrid. Entre 1977 y 1983 elaboró modelos numéricos en software para diversos cálculos en el Instituto Nacional de Energía Nuclear. En 1991 y 1992 colaboró en la escuela Superior de Ingenieros de Telecomunicaciones de la Universidad Politécnica de Madrid en la elaboración de prototipos para los codificadores de voz de telefonía celular. En 1993 y 1994 trabajó en la Secretaría de Hacienda en el prototipo del sistema automatizado de cálculo del ISR para personas morales. En ProceSar, contribuyó a la determinación de la cantidad de cuentas duplicadas en el sistema de cuentas de ahorro para el retiro. Ha trabajado en instituciones financieras y bancarias en la elaboración de modelos de detección de riesgos y fraudes, y en instituciones públicas, privadas y asociaciones de empresas en la realización de modelos y metodologías para compilar índices para estudiar la evolución de indicadores de interés. Desde 1983 es profesor de tiempo completo en el Departamento Académico de Computación de la División de Ingeniería del ITAM, donde además participó en el diseño de los planes de la carrera de Ingeniería en Computación. Como docente ha propuesto varias materias en el ámbito de la programación, el procesamiento de datos por medio de plataformas distribuidas y las aplicaciones de las tecnologías de información. Ha desempeñado cargos como funcionario académico en la División de Ingenierías del ITAM.