Syllabus de Preprocesamiento para Ciencia de Datos

BJETIVOS DEL CURSO: General: Entender el Proceso de la Ciencia de Datos y a mejorar la calidad de datos mediante el preprocesamiento de los datos para posteriormente generar modelos. **Específicos:** Conocer y aplicar los procesos de integración, exploración, perfilado, limpieza, de-duplicación, reducción y transformación de los datos.

PROGRAMA ANALÍTICO DEL CURSO

1. Calidad de datos
1.1 Introducción
1.1.1 Causas de baja calidad en lo datos
1.1.4 Interdependencia de dimensiones de calidad de datos
1.2 Modelo de Referencia de calidad de datos
1.3 Modelo de Medición de calidad de datos
1.4 Modelo de Evaluación de calidad de datos
1.5 Prácticas
2 El proceso de Ciencia de datos
2.1 Entendimiento del Negocio y especificación de la problemática
2.2 Heterogeneidad, Integración y Selección de datos
2.3 Exploración de Datos por estadística descriptiva
2.4 Perfilado de datos
2.5 Preparación de Datos (limpieza, reducción, transformación)
2.6 Modelado, evaluación, interpretación, generación de conocimiento, toma de decisiones, monitoreo
3. Integración de datos heterogéneos
3.1 Sistemas de bases de datos heterogéneos
3.1.1 Heterogeneidad semántica
3.1.2 Heterogeneidad sintáctica
3.2 Inconsistencias extensionales e intensionales
3.2.1 Mapeo y correspondencia de esquemas
3.3 Prácticas
4. Limpieza de datos
4.1 Proceso de de-duplicación de registros en base de datos
4.2 Funciones para fusión de registros
4.3 Obtención del registro de oro o único
4.4 Tipos de datos y contenidos
4.5 Corrección de datos incompletos
4.6 Corrección de datos faltantes
4.7 Prácticas
5. Reducción, Discretización y Transformación
5.1 Métodos de discretización
5.2 Métodos de Reemplazo
5.3 Métodos de Agrupación
5.4 Métodos de Reducción
5.5 Métodos de transformación
5.6 Prácticas
6. Minería de Datos
6.1 Metodología CRISP
6.2 Introducción a Minería de datos
6.3 Tareas descriptivas
6.4 Tareas predictivas
6.5 Distancias y similitudes
6.6 Correlación
6.7 Agrupamiento
6.8 Reglas de Asociación
6.9 Evaluación de modelos
6.10 Prácticas con Herramienta de minería de datos (p.e. Rapid Miner)
7 Aprendizaje Supervisado
7.1 Regresión, varianza,etc.
7.2 Arboles de decisión (ID3, C45)
7.3 Vecinos más cercanos
7.4 Naive Bayes
7.5 Máquinas de vectores de soporte
7.6 Introducción a redes neuronales
7.7 Redes neuronales profundas
7.8 Introducción a Aprendizaje por refuerzo

PROGRAMA GENERAL DE ACTIVIDADES DEL CURSO

Syllabus de Preprocesamiento para Ciencia de Datos

A continuación, se describen las actividades que se desarrollarán durante el semestre en la materia calidad y preprocesamiento de datos impartido por la Dra. Maria del Pilar Angeles.

Semana	Notas, presentación	Fechas	Temas
1	EvolucionSistInfo.pptx IntroCalidadDeDatos.pptx Calidad-DATOS-1.pptx EjemploAplicacionDataQuality.pptx	Ago. 12 y 14	Presentación, evaluación, 1.1 Introducción a la calidad de datos 1.1.1 Causas de problemas de calidad de bases de datos 1.1.2 Calidad de información 1.1.3 Calidad de datos 1.1.4 Interdependencia de dimensiones de calidad de datos 1.2 Modelo de Referencia de calidad de datos 1.3 Modelo de Medición de calidad de datos 1.4 Modelo de Evaluación de calidad de datos
2	Explicación proyecto de Calidad de datos Intro-ProcesoCienciaDatos.pptx ProblemaSelecyExploraDatos **Practica P0-CayPre-Instal**	Ago 19,**21**	Explicación y dudas proyecto de calidad de datos 2 El proceso de Ciencia de Datos Perfilado de datos **PRACTICA 0 de softwares a usar**
3	2 IntroDataProfiling Proceso-DataPreparation.pptx 3. IntegraciondeInformacion.pptx Heterogeneidad.pptx	Ago. 26**,** **28**	3. Integración de datos heterogéneos 3.1 Sistemas de bases de datos heterogéneos 3.1.1 Heterogeneidad semántica 3.1.2 Heterogeneidad sintáctica 3.2 Inconsistencias, mapeo de esquemas
4	**P1-CayPre-Extracción-e-Integra** 4 LimpiezadeDatos IntroDataMatching.pptx PreprocesamientoDataMatching-vcorta.pptx MetodosdeIndexado-vcorta.pptx MetodosdeComparacion-vcorta.ppts	*Sep **2**,**4***	***Primer examen parcial sobre Temas 1,2 y 3 (2)*** ***P*ráctica 1 extracción de datos e integración** 4 LimpiezadeDatos 4. Limpieza-deduplicación 4.1 Detección y medición de registros duplicados (unicidad)- Codificación, indexación, comparación, clasificación) ***Exposición de proyectos avance 1 (4)***
5	MetodosdeComparacion-vcorta.pptx MetodosdeClasificacion-vcorta.pptx Evaluaciondeclasificacionderegistros-vcorta.pptx **P2-CayPre-Dedup-Python-recordlinkage**	Sep. 9,***11***	4.1 Detección y medición de registros duplicados (unicidad)- Codificación, indexación, comparación, clasificación) 4.2 Corrección de datos incompletos **Práctica 2 detección duplicados (11)**
6	EjercicioDeduplicacion en clase TareaEjercicioDeduplicacion IntroFebrl.pptx, **P3-CayPre-Dedup-Eval**	Sep **18**	***Exposición, entrega avance 2*** Funciones para fusión de registros Obtención del registro de oro o único **Practica 3**
7	LimpiezaFusionDatos DetecciónyLimpiezadeDatosFaltantes LimpiezaDatosAnomalos-Ruido.pptx	Sep. ***23***,**25**	***Exposición entrega avance 3 (25)***
8	**P4-CayPre-Limpieza-Fusion** 5 ReduccionyDiscretización ReduccionyDiscretizaciondeDatos.pptx **P5-CayPre-Prepro**	Sep. 30, Oct. **2**	5. Reducción Discretización y transformación **Práctica 4 y 5 (2)**
9	5 TransformacionDatos.pptx **P6-CayPre-Preproces-ClasifPython**	Oct. 7,**9**	5. Reducción Discretización y transformación **Práctica 6 (9)**
10	6 MineriaDatos.pptx O AprendizajeMaquina-Mineria	Oct. 14,**16**	***Exposición entrega avance 4 limpieza según proyecto (16)*** Algoritmos de minería de datos para perfilado, preprocesamiento y transformación de datos**.**

	P7-CayPre-EvaluacionCasyRegrPython		
11	6 MineriaDatos.pptx	Oct. 21,23	Algoritmos de minería de datos para perfilado, preprocesamiento y transformación de datos.
12	6 Minería de datos	Oct. 28,**30**	Ejemplos de algoritmos de minería de datos. Aplicar técnicas de minería a los datos listos. Prácticas con Herramienta de minería de datos (p.e. Rapid Miner) ***Segundo Examen parcial sobre Temas 4 y 5***
13	Minería de datos y 7 Aprendizaje Supervisado	Nov. ***4,***6	7. Aprendizaje supervisado 7.1 Regresión lineal, varianza 7.2 Arboles de decisión (ID3, C45) ***Exposición entrega y exposición avance 5(4)***
14	7 Aprendizaje Supervisado	Nov 11,13	7.3 Vecinos más cercanos 7.4 Naive Bayes 7.5 Máquinas de vectores de soporte
15	7 Aprendizaje Supervisado	Nov. **18**,*20*	***Exposición entrega avance 6(18)***
16	Aclaración y revisión proyectos Calidad de Datos	Nov 25,**27**	**Tercer examen parcial: Tema 6** ***Exposición de proyectos entrega final con implementación de observaciones de las entregas anteriores y con los12 puntos del formato de entrega (27)***
	Aplicación examen final, calificación y revisión de exámenes	Dic 1	***1er. Examen Final*** Entrega calificación en página web
	Aplicación examen final, calificación y revisión de exámenes	Dic 8	***2do. Examen Final***, Entrega calificación en página web

EVALUACIÓN DEL CURSO

Evaluación de las actividades y el peso relativo de cada grupo de ellas para conformar la calificación final del curso.

Actividad	Porcentaje
Exámenes Parciales	50%
Participaciones y tareas	10%
Prácticas OBLIGATORIAS	20%
Proyecto OBLIGATORIO	20%
Total	100%
Examen Final	100%

Por reglamento general de exámenes, se tienen tres oportunidades para acreditar la materia:
1.- Presentar TODOS los elementos correspondientes a la evaluación del curso (tabla anterior) en tiempo y forma. Al obtener un promedio mayor o igual a 7.6 se da por acreditado el curso, no presenta final.
2.- Presentar el primer examen final (aborda todos los temas) y su calificación se pone en actas.
3.- Presentar el segundo examen final (aborda todos los temas) y su calificación se pone en actas.
Para las calificaciones con enteros de 6 en adelante y decimales .6 sube al siguiente entero. Ej. 5.6 = 5, 6.6=7.
Por reglamento general de exámenes no se puede presentar final para subir de calificación.
No se acredita la materia si se obtiene calificación menor o igual a 5.9 en exámenes finales.

BIBLIOGRAFIA

Básica

1. Loshin, D. (2011). The practitioner's guide to data quality improvement. Burlington, MA: Morgan Kaufmann.

2. Maydanchik, A. (2007). Data quality assessment. Bradley Beach (N.J.): Technics Publications.
3. McGilvray, D. (2008). Executing data quality projects. Amsterdam: Morgan Kaufmann/Elsevier.
4. Redman, T. (2001). Data quality. Boston: Digital Press.
5. Sebastian-Coleman, L. (2013). Measuring data quality for ongoing improvement. Amsterdam: Morgan Kaufmann.
6. West, M. (2011). Developing high quality data models. Burlington, MA: Morgan Kaufmann

Complementaria

1. Chan, C. (2009). Data quality and high-dimensional data analysis. Singapore: World Scientific.
Wang, R., Ziad, M., & Lee, Y. (2002). Data quality. London.: Kluwer Academic.

SOFTWARES:

-Pentaho Community Edition, contiene data quality, data cleansing y data integration:
https://github.com/ambientelivre/legacy-pentaho-ce?tab=readme-ov-file
o bien:
Registrarse en Qlik program academy https://www.qlik.com/us/company/academic-program para poder bajar Qlik Cloud, ahí hay cursos y un tenant con duración de un año. Sin embargo, se requieren para las prácticas: Qlik Talend Data Quality, Qlik Talend Data Cleansing y Qlik Talend Data Integration los tres contenidos en Qlik Talend Data Fabric: https://www.qlik.com/us/products/talend-data-fabric 14 dias solamente a partir de la instalación
Las prácticas pueden ser realizadas con Qlik Talend o con Pentaho
-Librería recordlinkage a instalar desde Python

PAGINA DE LA MATERIA:
http://lcd.iimas.unam.mx/profesores/pilarang

PRÁCTICAS:

Las prácticas se entregan vía correo electrónico dirigido a temasselectosbd@yahoo.com.mx **el día de la fecha de entrega antes de las 2 de la tarde** los entregables consisten en lo siguiente:
1.- El Asunto del correo electrónico debe ser el **número, tipo de práctica y el equipo,** (ej. P1-CayPre-Integra, Equipo 3), en el cuerpo del correo deben estar los nombres de los participantes.
2.- El Archivo que contenga las imágenes de las pantallas en donde se reflejen la ejecución de los comandos completos que se requieren para realizar todos los pasos de todas las actividades y sus correspondientes resultados.
3.- Los **archivos deben estar adjuntos al correo electrónico,** NO deben ser parte del texto del mensaje.

NOTAS:

a) Los correos deben tener en el Asunto el número de práctica y el número de Equipo, en el cuerpo del correo deben estar los nombres completos de todos los integrantes que participaron en la elaboración de la práctica.
b) En caso de que no se envíe a la dirección indicada, fuera de horario establecido o sin los archivos adjuntos correspondientes, se considera la práctica como no entregada. LO MISMO APLICA PARA EL PROYECTO.

Fechas de entrega de prácticas

Practica P0-CayPre-Instal	21 agosto
Practica P1-CayPre-Integra	4 septiembre
Practica P2-CayPre-DetDup-Python-recordlinkage	11 septiembre
Práctica P3-CayPre-Perfilado-Tal	18 septiembre
Práctica P4-CayPre-Limpieza-Fusion	2 octubre
Práctica P5-CayPre-Prepro	2 octubre
Práctica P6-CayPre-ClasificarPython	9 octubre
Práctica P7-CayPre-ClasyRegPython	16 octubre

PROYECTO:

El proyecto completo debe entregarse en papel con hojas numeradas y en formato electrónico por correo electrónico a la misma dirección que las prácticas, éste debe contener al igual que las prácticas, código en formato txt y pantallas con la ejecución correcta de las instrucciones.

Se requiere entregar proyecto para tener derecho al primer final. Si no se entrega proyecto, se presentará solo el segundo final.

NINGUNA FECHA PROGRAMADA SUFRE MODIFICACION.

A continuación, se especifica el proyecto, su formato de entrega final y las entregas parciales

Descripción de avances en el proyecto de calidad de datos

Avance 1 Encontrar bases de datos participantes y generar una base de datos heterogénea federada en español en el sector Salud.

Avance 2 Problemática y entendimiento del negocio dentro del sector salud con base a la base de datos heterogéneas.

Avance 3 Extracción y perfilado de datos relevantes

Avance 4 Limpieza de datos relevantes

Avance 5 Generación de consultas heterogéneas con LLM

Avance 6 Análisis de datos a través de minería de datos descriptiva, prescriptiva o predictiva

Entrega Final: Los 12 puntos establecidos en el Formato de entrega.

La entrega final del proyecto contendrá todos los puntos revisados y mejorados durante la revisión de cada avance.