ITESO
Universidad Jesuita de Guadalajara

Departamento de Matemáticas y Física
Guía de Aprendizaje: Laboratorio de Procesamiento de Datos

Asignatura:	**Laboratorio de Procesamiento de Datos**			**Créditos BCD:**	**4**
Clave:	**MAF3658A**	**Grupo:**		**Créditos TIE:**	**4**
Horario:	**7:00 – 9:00**	**Carreras:**	**Ingeniería y Ciencia de Datos**	**Salón:**	
Departamento:	**Matemáticas y Física**	**UAB:**		**Periodo:**	**Primavera 2022**
Coordinador UAB:	**Juan Diego Sánchez Torres**		**E-mail:**	**dsanchez@iteso.mx**	**Ext. 3069**
Profesor:	**Raúl Romero Barragán**		**E-mail:**	**raul.romero@iteso.mx**	**NA**

BCD: *Bajo Conducción Docente* TIE: *Trabajo Independiente del Estudiante*

Presentación

En cualquier ámbito profesional existe la constante necesidad de mejorar los procesos internos, ser más eficiente hacia y para los clientes además de poder encontrar nuevas metodologías que permitan llegar a mercados distintos. Existe una serie de herramientas medulares que muchas empresas generan a carretadas y pocas de ellas las notan (aún en el siglo XXI), menos aún las utilizan; éstas han sido conocidas por muchos como el *nuevo petróleo* haciendo referencia a lo útil y valioso que dichos recursos son: **los datos.**

Cualquier empresa, sea del rubro que sea, tiene la capacidad de generar datos. Los datos tienen toda la información que la misma compañía se ha encargado de generar y con ello tiene el conocimiento no solo de sus transacciones compra-venta, sino también del tipo de cliente que tiene (que puede ser el cliente al que se buscaba llegar en un inicio o no), de la fidelidad del cliente (qué tanto un cliente continúa interactuando con el ecosistema de la empresa), de las necesidades tanto propias como las que el mercado le demande entre muchas más.

El problema al que se enfrenta cualquier ente (sea empresa pública o privada) que haya descubierto el potencial de sus propios datos, radica en que muchas veces éstos carecen de orden, estructura y a veces incluso lógica. ¿De qué va a servir un coche lujoso sin el combustible que lo haga explotar todas sus capacidades? ¿De qué sirve tener una computadora con las mejores características en el mercado sin que ésta cuente con una fuente de alimentación?

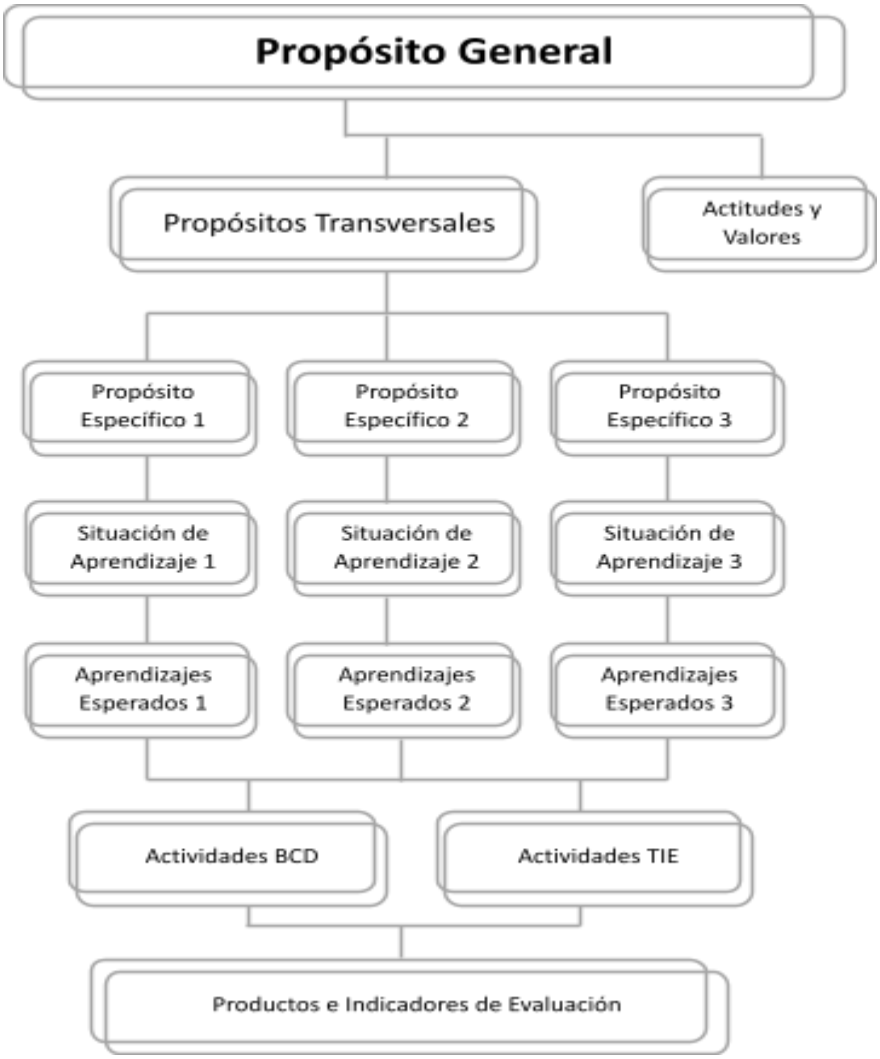
La asignatura *Laboratorio de Procesamiento de Datos* presenta técnicas de lectura, estructura y sobre todo tratamiento y/o selección de información tal que ésta pueda ser analizada y sobre todo explotada por entes públicos y/o privados para el aprovechamiento de la misma. Los puntos anteriormente expuestos serán realizados mediante lenguajes de programación, técnicas de análisis estadístico así como de ML (*machine learning*).

Contextualización de la materia

_____Para poder lograr un desempeño satisfactorio en la asignatura es de vital importancia contar con conocimiento elementales así como gusto por la programación en Python. No es necesario (pues será desarrollado de manera indirecta durante el curso) pero sí deseable el uso de herramientas estadísticas básicas para el análisis masivo de información.

Al finalizar con éxito la asignatura se contará con herramientas básicas para extraer información de diferentes fuentes, herramientas para la compresión y posterior limpieza de la información además de una serie de técnicas que permitan transformar y construir nuevos datos a partir de los existentes.

Mapa descriptivo



Propósitos de la materia

General

Analizar y darle estructura a múltiples bases de datos para generar un *dataset* con las suficientes herramientas que requerirá un proceso de modelación de ML (*machine learning*).

Transversales

1. Colaboración y fomento de debate sobre el uso y significado de la información con compañeros.
2. Generar el hábito de resolver problemas con herramientas computacionales.
3. Usar de manera efectiva y eficaz las tecnologías de información para representar e interpretar los conceptos en diferentes formas: numérica y algorítmica.
4. Implementar en Python las soluciones de análisis y tratamiento de información.
5. Desarrollo de la creatividad y búsqueda de información.

Específicos

1. Extraer y ordenar datos a partir de diversas fuentes y formatos.
2. Desarrollar la capacidad de hacer un análisis descriptivo de información.
3. Identificación de errores, faltas e inconsistencias en los datos así como formas de tratarlos.
4. Crear nueva información a partir de la ya existente.
5. Presentar un *dataset* organizado tal que pueda cumplir un objetivo posterior de modelación y predicción.

Actitudes y valores

Se espera que al cursar esta asignatura se desarrolle responsabilidad ante la actividad académica, manifiesta en al menos los siguientes aspectos:

1. Participación activa, con compromiso, perseverancia y actitud positiva.
2. El cumplimiento de las normas de disciplina establecidas.
3. El cumplimiento en tiempo y forma de las actividades que se encomienden como trabajo independiente.
4. El desarrollo de espíritu crítico y autocrítico (constructivo) en el análisis del desempeño propio y de los compañeros.
5. El sentido de la ética, evitando, en particular, cometer actos deshonestos en la realización de las actividades evaluativas.
6. El desarrollo de la capacidad para identificar características personales al afrontar procesos de aprendizaje y, como consecuencia, para aprender con mayor independencia.
7. Diálogo abierto, directo y respetuoso tanto con el profesor como con los compañeros.
8. Tolerancia y respeto.

Evaluación del aprendizaje

Productos	Porcentaje de calificación
Tareas	40%
1. Proyectos de aplicación a. Extracción de datos b. Comprensión de datos c. Limpieza y transformación d. *Encoding* y construcción de variables a partir de datos categóricos	1. 30% a. 5% b. 5% c. 10% d. 10%
Examen final	30%
TOTAL:	**100%**

1. Tareas
    a. En cada sesión habrá una tarea salvo indicación expresa del profesor.
2. Proyectos de aplicación
    a. Al final de cada módulo existirá un proyecto.
    b. Cada proyecto tendrá sus propios lineamientos.
    c. Los lineamientos de cada proyecto se darán a conocer conforme avance el curso.
3. Exámenes
    a. El examen será individual y *para llevar*. Ésto significa que el examen durará **una semana** abierto en la plataforma pero el tiempo de realización una vez abierto será de **12 horas**.
    b. Los exámenes contendrán una parte teórica y una parte práctica.
    c. Para más detalles sobre la forma de calificar consultar el archivo “Rúbricas de evaluación”.

Políticas y lineamientos del curso

Tiempo aproximado del curso

1. Familiarización con herramientas básicas como *pandas, numpy* y *notebooks* (0.5 semanas).
2. Extracción de distintas fuentes de información (1.5 semanas).
3. Comprensión de datos (4 semanas).
4. Limpieza y transformación de datos (5 semanas).
5. *Encoding* y construcción de variables a partir de datos categóricos (2 semanas).
6. Selección de características (2 semanas).

Lineamientos del curso

1. La plataforma oficial del curso es Canvas.
2. El lenguaje que se usa en clase, tareas, entregas y demás es únicamente python >= 3.5.
3. Los medios de comunicación oficiales fuera de clase son la plataforma y *Slack*.

    a. Una vez iniciado el curso, el profesor dará de alta a todos los alumnos y a él mismo en la plataforma *Slack*. Este será un medio en el cual se podrán preguntar dudas y mantener debate sobre los temas de clase.
4. Es responsabilidad del alumno mantenerse al día de los avisos relativos a la materia.
    a. Los avisos se darán primariamente por *Slack* y posteriormente se pondrán en la plataforma.
5. A menos que la actividad lo requiera y únicamente bajo indicación del profesor, **no se permite el uso de celulares**.
    a. Si el alumno requiere hacer una llamada por asuntos personales tendrá que salir del salón.
    b. El uso constante del celular durante clases será acreedor hasta a 3 llamados de atención. Al tercer llamado de atención por este tema, el alumno se hará acreedor a una falta.
6. **Bajo ninguna circunstancia se permite la deshonestidad académica**. Cualquier trabajo que tenga evidencia de copia será anulación automática del trabajo tanto para el alumno que copia como para el que accede a que su trabajo sea copiado.
7. En caso de asistencia presencial al campus, el uso de cubrebocas será obligatorio durante la sesión.
8. En caso de clase en línea, es requisito de clase tener su cámara encendida.
9. Durante cada módulo existirán *quizzes* que estarán en la plataforma y solo podrán ser realizados una vez bajo un tiempo predefinido.
10. Al finalizar cada módulo habrá un proyecto de aplicación que podrá consistir de una exposición o un ejercicio reto. Se comunicará con anticipación en qué consistirá cada uno de estos proyectos así como los lineamientos específicos de los mismos.
11. Cada sesión habrá ejercicios (a menos que el profesor indique lo contrario) que el alumno deberá de completar de manera autónoma **y subirlos a la plataforma** con una fecha establecida.
12. Todas las tareas, trabajos y demás serán entregados mediante la plataforma en la hora y fecha indicadas, **no se recibirán trabajos a destiempo** ni mediante otro medio que no sea la plataforma oficial.
13. El horario de las clases inicia a las 7:00, se tomará lista a las 7:10 de la mañana y con eso se dará comienzo a la clase. Alumno que no esté presente en dicha hora o no conteste por la razón que sea será acreedor a una falta en la sesión.
14. En clase no se permite hacer desorden que afecte la impartición de la misma.
    a. Se harán hasta 3 llamados de atención por el tema, al tercer llamado de atención el alumno se hará acreedor a una falta.
15. Se permite como máximo el **20%** de inasistencias al curso, equivalente a **6 inasistencias** para mantener el derecho a calificación.
16. Se requiere que el promedio de la calificación de los proyectos de aplicación en el curso sea aprobatorio para tener derecho a calificación.
17. Habiendo cumplido el criterio anterior, la calificación aprobatoria es **mayor o igual a 6.0** (sesenta por ciento) sin aplicar criterios de redondeo si se encuentra por debajo de este límite.

Referencias

El profesor entregará documentos de estudio y/o ampliará las fuentes/referencias aquí presentes con información relevante durante el desarrollo del curso vía la plataforma oficial.

1. Python Business Intelligence Cookbook; R. Dempsey, 2015, ISBN: 978-1-78528-966-8.
2. An Introduction to Machine Learning, Kubat, M., Springer, 2015, ISBN: 978-3-319-20009-5.
3. Ligas de papers/conferencias interesantes relacionados a ML:
    a. https://arxiv.org/list/stat.ML/recent
    b. https://proceedings.neurips.cc/paper/2020
    c. https://www.guide2research.com/topconf/machine-learning
    d. https://sigmod2018.org/
    e. https://blog.acolyer.org/
    f. Algunos post de medium/analyticsvidhya (se irán recomendando durante el curso).

Fuentes de apoyo y datos

1. https://datos.jalisco.gob.mx/
2. http://www.beta.inegi.org.mx/datos/
3. https://www.kaggle.com/datasets
4. http://www.jaliscocomovamos.org/datos-abiertos
5. https://www.datos.gob.mx/
6. https://www.kdnuggets.com/datasets/index.html
7. https://finance.yahoo.com/
8. https://mybinder.org/v2/gh/choldgraf/nbreport/master?filepath=example%2Fan_example_notebook.ipynb