

### PROGRAMA DE ESTUDIOS

#### NOMBRE DE LA ASIGNATURA

Introducción a la ciencia de datos

SEMESTRE	CLAVE DE LA ASIGNATURA	TOTAL DE HORAS
Octavo semestre	075083	80

#### OBJETIVO(S) GENERAL(ES) DE LA ASIGNATURA

Que el alumno adquiera los conocimientos básicos necesarios en ciencia de datos, conozca los diferentes tipos de bases de datos, sea capaz de recopilar, limpiar y manejar grandes cantidades de datos, que comprenda la forma de visualizar los resultados, que reconozca los modelos utilizados en el proceso de minería y los distintos métodos para evaluarlos.

#### TEMAS Y SUBTEMAS

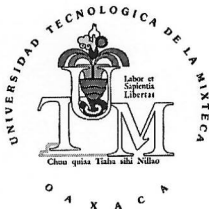
1. **¿Qué es la ciencia de datos?**
  - 1.1. Informática, ciencia de datos y ciencia real.
  - 1.2. Revisión de ejemplos reales.
  - 1.3. La base de datos de películas de Internet (*IMDb*).
  - 1.4. *Ngramas* de Google.
  - 1.5. Propiedades de los datos.
  - 1.6. Datos estructurados frente a datos no estructurados.
  - 1.7. Datos cuantitativos contra categóricos.
  - 1.8. Big Data vs. Little Data.
  - 1.9. Clasificación y regresión
2. **Gestión de datos**
  - 2.1. Lenguajes para la ciencia de datos.
  - 2.2. Entornos portátiles (*Notebooks*).
  - 2.3. Formatos de datos estándar.
  - 2.4. Recopilación de datos.
  - 2.5. Caza.
  - 2.6. Raspado.
  - 2.7. Registro.
  - 2.8. Limpieza de datos.
  - 2.9. Errores frente a artefactos.
  - 2.10. Compatibilidad de datos.
  - 2.11. Manejo de valores perdidos.
  - 2.12. Detección de valores atípicos.
  - 2.13. Extracción del conocimiento de multitudes (*Crowdsourcing*).
  - 2.14. Definición.
  - 2.15. Mecanismos de agregación.
  - 2.16. Servicios de *Crowdsourcing*.
  - 2.17. Ludificación.
3. **Puntuaciones y clasificaciones (Scores and Rankings)**
  - 3.1. Extracción de características.
  - 3.2. El índice de masa corporal (IMC).
  - 3.3. Desarrollo de sistemas de puntuación (scores).
  - 3.4. Estándares de oro y proxies.
  - 3.5. Scores vs. Rankings.
  - 3.6. Reconocimiento de buenas funciones de puntuación.
  - 3.7. Z-score y normalización.
  - 3.8. Técnicas avanzadas de clasificación.



**PROGRAMA DE ESTUDIOS**

- 3.9. Clasificaciones Elo.
- 3.10. Combinación de clasificaciones
- 3.11. Clasificaciones basadas en dígrafos.
- 3.12. PageRank.
- 3.13. Teorema de imposibilidad de Arrow.
- 4. Visualización de datos**
  - 4.1. Análisis exploratorio de datos.
  - 4.2. Análisis estadístico de un conjunto de datos.
  - 4.3. Estadísticas y Cuarteto de Anscombe.
  - 4.4. Herramientas de visualización.
  - 4.5. Desarrollo de una visualización estética.
  - 4.6. Maximización de la relación datos-tinta.
  - 4.7. Minimizar el factor de mentira.
  - 4.8. Minimizar *Chartjunk*.
  - 4.9. Escalado y etiquetado adecuados.
  - 4.10. Uso efectivo del color y el sombreado.
  - 4.11. El poder de la repetición.
  - 4.12. Grandes visualizaciones.
  - 4.13. Horario de trenes de Marey.
  - 4.14. Mapa de cólera de Snow.
  - 4.15. Año meteorológico de Nueva York.
  - 4.16. Lectura de gráficos.
  - 4.17. La distribución oculta.
  - 4.18. Sobreinterpretación de la varianza.
  - 4.19. Visualización interactiva.
- 5. Modelos matemáticos**
  - 5.1. Filosofías del modelado.
  - 5.2. Navaja de Occam.
  - 5.3. Compensaciones entre sesgo y varianza (Bias-variance).
  - 5.4. Manejo ético de datos.
  - 5.5. Una taxonomía de modelos.
  - 5.6. Modelos lineales contra no lineales.
  - 5.7. Cajas negras contra modelos descriptivos.
  - 5.8. Modelos basados en datos y de primer principio.
  - 5.9. Modelos estocásticos contra deterministas.
  - 5.10. Modelos planos frente a modelos jerárquicos.
  - 5.11. Modelos de referencia.
  - 5.12. Modelos de referencia para la clasificación.
  - 5.13. Modelos de referencia para la predicción de valor.
  - 5.14. Evaluación de modelos.
  - 5.15. Evaluación de clasificadores.
  - 5.16. Curvas de características de receptor-operador (ROC).
  - 5.17. Evaluación de sistemas multiclase.
  - 5.18. Evaluación de modelos de predicción de valor.
  - 5.19. Entornos de evaluación.
  - 5.20. Higiene de los datos para la evaluación.
  - 5.21. Amplificación de pequeños conjuntos de evaluación.
  - 5.22. Modelos de simulación.





### PROGRAMA DE ESTUDIOS

#### ACTIVIDADES DE APRENDIZAJE

Sesiones dirigidas por el profesor en las que presenta los conceptos y, al mismo tiempo, se realizarán programas que ilustran cada uno de los conceptos, se sugiere utilizar algún Notebook como Collaboratory o Jupyter para realizar programas con el lenguaje Python, así como Kaggle y GitHub para compartir y descargar algoritmos programables. Se recomienda ampliamente impartir el curso en un laboratorio con equipo de cómputo disponible para cada estudiante. El contenido será visto de forma panorámica y se iniciará con un proyecto desde el inicio del curso con fines didácticos y de práctica de cada uno de los temas.

#### CRITERIOS Y PROCEDIMIENTOS DE EVALUACIÓN Y ACREDITACIÓN

En términos de los artículos 25 incisos (b), (e), (f) y (g); del 48 al 62, del Reglamento de alumnos de licenciatura aprobado por el H. Consejo Académico el 19 de mayo del 2016, los lineamientos que habrán de observarse en lo relativo a los criterios y procedimientos de evaluación y acreditación, entre lo más importante:

Al inicio del curso el profesor deberá indicar el procedimiento de evaluación que deberá comprender, al menos tres evaluaciones parciales que tendrán una equivalencia del 50% de la calificación final y un examen ordinario que equivaldrá al restante 50%.

Las evaluaciones podrán ser escritas y/o prácticas y cada una consta de un examen teórico-práctico, tareas y proyectos. La parte práctica de cada evaluación deberá estar relacionada con la ejecución exitosa y la documentación de la solución de problemas sobre temas del curso.

Además, pueden ser consideradas otras actividades como: el trabajo extra-clase, la participación durante las sesiones del curso y la asistencia a las asesorías.

El examen tendrá un valor mínimo de 50%; las tareas, proyectos y otras actividades, un valor máximo de 50%.

#### BIBLIOGRAFÍA (TIPO, TÍTULO, AUTOR, EDITORIAL Y AÑO)

##### Libros Básicos:

1. **Data science design manual**, Steven S. Skiena, Text in computer science, Springer 2017.
2. **Introducing data science, big data, machine learning and more using Python tools**, Davy Cielen, Arno D. B. Meysman and Mohamed Ali, Manning Publications Co., 2016.
3. **Data Science**, Lillian Pierson, John Wiley & sons, 2015.

##### Libros de Consulta:

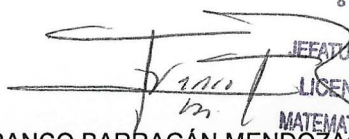
1. **Python for data Science**, Luca Massaron and John Paul Mueller, John Wiley & sons, 2015.
2. **Introducción a la minería de datos**, José Hernandez Orallo, Ma. José Ramírez Quintana, Cesar Ferri Ramirez, Pearson, Prentice Hall.
3. **Data Science strategy**, Ulrika Jägare, John Wiley & sons, 2019.

#### PERFIL PROFESIONAL DEL DOCENTE

Estudios de maestría o doctorado en matemáticas, matemáticas aplicadas o computación con conocimientos en ciencia de datos y machine learning.

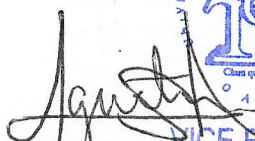
Vo.Bo.



  
JEFE DE CARRERA  
LICENCIATURA EN  
MATEMÁTICAS APLICADAS  
DR. FRANCO BARRAGÁN MENDOZA  
JEFE DE CARRERA

AUTORIZÓ



  
VICE-RECTORIA  
VICE-RECTOR ACADÉMICO  
DR. AGUSTÍN SANTIAGO ALVARADO  
VICE-RECTOR ACADÉMICO