Ciencia de Datos y Machine Learning

Contents

BIENVENIDA **5**

Objetivo 5

Alcances del curso 5

Instructores 5

Duración y evaluación del curso 10

Recursos y dinámica de clase 11

BIENVENIDA

Objetivo

Capacitar y brindar acompañamiento al equipo de analítica de ASERTA en temas relacionados con Ciencia de Datos para la correcta implementación de proyectos y toma de decisiones basadas en la evidencia de datos internos y externos a la empresa para lograr un beneficio operativo y económico.

Alcances del curso

El participante conocerá los conceptos teóricos alrededor de esta ciencia y sabrá implementar correctamente un análisis exploratorio estadístico y gráfico que le permita conocer a mayor profundidad los datos a usar. Conocerá y sabrá implementar los modelos predictivos de Machine Learning más usados y de mayor impacto en la industria de seguros y fianzas. Finalmente, sabrá tomar decisiones sobre el correcto uso e implementación de los modelos para aumentar el beneficio comercial dentro de la institución.

Instructores

ACT. ARTURO BRINGAS

LinkedIn: arturo-bringas **Email:** act.arturo.b@ciencias.unam.mx

Actuario egresado de la Facultad de Ciencias y Maestría en Ciencia de Datos por el ITAM. Se especializa en modelos predictivos y de clasificación de *machine learning* aplicado a seguros, marketing, deportes y movilidad internacional. Es jefe de departamento en Investigación Aplicada y Opinión de la UNAM, donde realiza estudios estadísticos de impacto social. Ha sido consultor *Sinior Data Scientist* para empresas y organizaciones como GNP, El Universal, UNAM, Sinnia, Geekend, la Organización de las Naciones Unidas Contra la Droga y el Delito (UNODC), entre otros. Actualmente es profesor de *Ciencia de datos y*

Machine Learning en AMAT y se desempeña como consultor independiente en diferentes proyectos contribuyendo a empresas en temas de análisis estadístico y ciencia de datos con machine learning.

ACT. KARINA LIZETTE GAMBOA

LinkedIn: KaLizzyGam **Email:** lizzygamboa@ciencias.unam.mx

Actuaria egresada de la Facultad de Ciencias, por la UNAM y candidata a Maestra en Ciencia de Datos por el ITAM.

Experiencia en áreas de analítica predictiva e inteligencia del negocio. Lead y Senior Data Scientist en consultoría en diferentes sectores como tecnología, asegurador, financiero y bancario. Es experta en entendimiento de negocio para la correcta implementación de algoritmos de inteligencia y explotación de datos. Actualmente se desarrolla como Arquitecta de Soluciones Analíticas en Merama, startup mexicana clasificada como uno de los nuevos unicornios de Latinoamérica. Senior Data Science en CLOSTER y como profesora del

diplomado de Metodología de la Investigación Social por la UNAM así como instructora de cursos de Ciencia de Datos en AMAT.

Empresas anteriores: GNP, Activer Banco y Casa de Bolsa, PlayCity Casinos, RakenDataGroup Consulting, entre otros.

Temario:

Módulo 1: Introducción a R (22 hrs)

Objetivo: A través de este módulo se adquirirán los conocimientos necesarios para la operación del software estadístico y la manipulación ágil de datos. Al finalizar, el participante desarrollará análisis exploratorios y reportes automatizados.

- Estructuras de almacenamiento de datos
  - Almacenamiento
  - Vectores
  - Matrices

  - Listas
  - DataFrames

- Funciones y estructuras de control

  - Librerías y funciones
  - Condicionamiento
  - Ciclos

- Manipulación de estructuras de datos

  - Importación de tablas
  - Consultas y transformación de estructuras
  - Iteraciones
  - Manipulación de texto y datos temporales

- Análisis exploratorio y visualización de datos

  - Guía de visualización
  - Análisis Exploratorio de Datos (EDA)
  - Análisis Gráfico Exploratorio de Datos (GEDA)
  - Reportes con markdown

- Consultoría y aplicaciones con datos institucionales

Módulo 2: Introducción a Ciencia de Datos (18 hrs)

Objetivo: Este módulo presenta los conceptos teóricos clave para conocer los términos, objetivo y alcances de proyectos con enfoque en ciencia de datos. Se presenta el flujo de trabajo y organización que deberá seguir un equipo para obtener el mayor beneficio posible. Adicionalmente, se propone presentar el software git y github para implementar correctamente el trabajo en equipo que garantice la reproducibilidad y seguridad del desarrollo realizado.

- Introducción a ciencia de datos

  - ¿Qué es la ciencia de datos?
  - Objetivo de ciencia de datos
  - Requisitos y aplicaciones
  - Tipos de algoritmos
  - Ciclo de vida de un proyecto
  - Taller de scoping
  - Perfiles de un equipo de ciencia de datos

- Concepto de Ciencia de Datos

  - Machine learning

  - Análisis supervisado
  - Análisis no supervisado
  - Sesgo y varianza
  - Pre-procesamiento e ingeniería de datos
  - Partición de datos

- Colaboración y reproducibilidad

  - Git & Github
  - Ambiente de desarrollo

- Consultoría y aplicaciones con datos institucionales

Módulo 3: Machine Learning: Supervisado (38 hrs)

Objetivo: Este módulo está diseñado para adquirir los conocimientos técnicos para conocer e implementar los distintos modelos de aprendizaje supervisado que son aplicados en ciencia de datos a la industria de los seguros y fianzas.

- Modelos de aprendizaje Supervisado

  - Regresión Lineal
  - Regresión logística
  - Regularización Ridge & Lasso
  - Elasticnet
  - KNN
  - Árbol de decisión
  - Bagging
  - Random Forest
  - Boosting
  - Stacking

- Toma de decisiones enfocadas a negocio

  - Comparación de modelos
  - Balance entre sesgo y cobertura
  - Cuantificación de sesgo e inequidad
  - Cuantificación de ganancia comercial
  - Diseño experimental

- Consultoría y aplicaciones con datos institucionales

Módulo 4: Machine Learning: No Supervisado

Objetivo: Este módulo permite al participante conocer técnicas de clustering para clasificar clientes de acuerdo con la utilidad y riesgo para la empresa. Adicionalmente, se presentan aplicaciones de clustering enfocadas a la estratificación de acuerdo con el riesgo geográfico.

- Técnicas de reducción de dimensión
  - Análisis de componentes principales
  - Creación de índices
- Clustering
  - Liga simple, compleja y promedio
  - Dendogramas & heatmaps
  - Kmeans &
  - Kmedoids
  - DBSCAN
- Consultoría y aplicaciones con datos institucionales

Requisitos:

Computadora con al menos 4Gb Ram.

Instalación de R con versión >= 4.1.0

Instalación de Rstudio con versión >= 1.4.17

Conocimientos generales de probabilidad, estadística y álgebra lineal

Duración y evaluación del curso

- El programa tiene una duración de 90 hrs.
- Las clases serán impartidas los días lunes a viernes, de 7:00 am a 9:00 pm
- Serán asignados ejercicios que el participante deberá resolver entre una semana y otra.
- Al final del curso se solicitará un proyecto final, el cual **deberá ser entregado para ser acreedor a la constancia de participación**.

Recursos y dinámica de clase

En esta clase estaremos usando:

- R (descargar)
- RStudio (descargar)
- Zoom Clases
  - **Pulgar arriba:** Voy bien, estoy entendiendo!
  - **Pulgar abajo:** Eso no quedó muy claro
  - **Mano arriba:** Quiero participar/preguntar ó Ya estoy listo para iniciar
- Google Drive
- Notas de clase
- Finalmente, se dejarán ejercicios que serán clave para el éxito del aprendizaje de los capítulos, por lo que se trabajará en equipo para lograr adquirir el mayor aprendizaje.