Educación Continua

TECNOLÓGICO DE MONTERREY

Diplomado

Data Science and AI: Del Concepto a Desarrollo de Aplicaciones - Live

Duración 120 horas

Objetivo general:
Aplicar las herramientas tecnológicas y matemáticas para desarrollar modelos de Ciencia de Datos para el perfil Data Science Jr, requeridos en organización/innovación/emprendimiento para la toma decisiones basadas en datos.

Beneficio:
Retroalimentación asíncrona a cada uno de los retos por parte de un experto.
Contar con personal capacitado para soportar una estrategia organización de tomar decisiones basada en datos.

Dirigido a:
(Analista, gerente como Científico de Datos, Jr)., Personal encargado de generar análisis descriptivos, inferenciales, predictivos y prescriptivos; dashboards para la toma de decisiones; desarrollar una aplicación de datos; desarrollar un emprendimiento.
Trabaja en una organización analizando datos en Tableau o equivalente; en sus estudios de licenciatura llevó cursos de algún lenguaje de programación de alto nivel; Trabajo en equipo, habilidades de comunicación.

Requisito:
Tener conocimiento de algún lenguaje de programación, y conocimientos básicos de probabilidad y estadísticas. Conocimiento en Excel. Realizar el test de perfilamiento.

Contenido:

Módulo 1 Fundamentos de Python

Diseñar programas en lenguaje de programación python sobre el ambiente de programación notebook, para que se cumpla con los requerimientos de la aplicación de Ciencia de Datos .

Temario

1. Introducción
2. Estándares en Python.
3. Estructuras de datos: lista, diccionarios, tuple, strings;
4. Estatutos condicionales.
5. Control de flujo
6. Definición de funciones
7. Manjeo de archivos csv. Incluir Identificar archivos vacíos. Crear una lista de archivos a procesar.
8. Ambiente de programación notebook (Jupiter y/o ?Google Colab).

Duración del módulo: 10 horas

Módulo 2 Plataformas de aplicación sobre Python

Diseñar programas en las plataformas de Panda y Numpy de python sobre el ambiente de programación notebook, para que se cumpla con los requerimientos de la aplicación de Ciencia de Datos requerida.

Temario

1. Introducción (Plataformas / Frameworks)
2. Manipulación de datos en Panda:
3. Manipulación de datos en Numpy
4. Instalar bibliotecas que no estén la pre-instaladas (Plataformas / Frameworks)

Duración del módulo: 10 horas

Módulo 3 Manipulación de datos en Python

Manipular una base de datos en la plataforma de Panda de python que involucre llamadas (queries) a las bases de datos y estatuos para concatenar y unir bases de datos.

Temario

1. Introducción
2. Algebra relacional
3. Funciones: join, merge, append
4. Ejemplo

Duración del módulo: 10 horas

Módulo 4 Plataformas de visualización

Explotar las ventajas y desventajas que tienen Python, y las plataformas de visualización matplotlib y seaborn, para generar gráficas que compartan ejes horizontal (x); el eje vertical (y) puede ser compartido a la misma escala, para que se cumpla con los requerimientos de interfase de la visualización requerida.

Temario

1. Introducción
2. Validación estadística del modelo
3. Plataformas de visualización: panda, matplotlib, y seaborn.
4. Gráficas superpuestas compartiendo eje horizontal (x) y vertical (y).
5. Gráficas superpuestas compartiendo eje horizontal (x) y con ejes verticales diferentes (y1, y2), cada uno con escala diferente.
6. Sub-gráficas (subplots)
7. Anotaciones de texto en las gráficas
8. Diferentes tipos de gráfias para exploración de las Datos: boxplot, distribuciones, dispersiones (Scatterplot), matriz de dispersiones (Scatterplot Matrix).

Duración del módulo: 10 horas

Módulo 5 Aplicación Web de Ciencia de Datos

Seleccionar las características de las bases de datos más importantes por su importancia para la visualización o por su importancia en mejorar la exactitud, precisión de un modelo de datos.

Temario

1. Imputación de valores perdidos.

2. Valores Atípicos (Handling Outliers)
3. Agrupamiento en contenedores (Binning)
4. Transformación Logarítmica
5. Distribución de valores de una columna en múltiples columnas binarias (One-Hot Encoding).
6. Operaciones en Grupos (groupby)
7. Separación de Datos (Feature Split)
8. Escalamiento (Scaling)
9. Extracción de fecha (Date)
10. Reducción de Dimensiones (FA, PCA, IDA)

Duración del módulo: 10 horas

Módulo 6 Ingeniería de Características

Desarrollar un Dashboard en la plataforma Streamlit que visualice datos interactivos para la visualización de datos requeridos por la organización para presentar KPIs o para toma de decisiones.

Temario

1. Fundamentos de UX/UI
2. Plataforma Streamlit
3. Manipulación de datos
4. Sidebar
5. Estatutos slider, radio, selectbox
6. Estatuto Caching

Duración del módulo: 10 horas

Módulo 7 Aprendizaje no-supervisado

Desarrollar modelo inteligentes no supervisado de datos utilizando scikit-learn de Python, seleccionando el número de agrupaciones adecuadas y analizando la efectividad del modelo utilizando medidas de calidad, que cumplan lo mejor posible con los requerimientos de la tarea requerida.

Temario

1. Introducción
2. Medidas de calidad de los algoritmos de clusterización.
3. Número de agrupaciones utilizando la gráfica K-Elbow.
4. Número de agrupaciones utilizando los valores de los coeficientes de Silhouette (gráfica Silhouette Plot).
5. Distancia relativa entre agrupaciones; importancia relativa de las agrupaciones.
7. Índice Calinski-Harabasz.

Duración del módulo: 10 horas

Módulo 8 Aprendizaje supervisado

Desarrollar modelos inteligentes supervisado de datos utilizando scikit-learn de Python; seleccionar el modelo adecuado y analizar la exactitud, precisión del modelo, que cumplan lo mejor posible con los requerimientos de la tarea requerida.

Temario

1. Introducción

2. Modelación de datos clásica: Árboles, Bosques, Gaussian Naive-Bayes, Regresión Lógica, Support Vector Machine.
3. Evaluación de modelos medida a utilizar. Matriz de Confusión, exactitud, precisión, F1, curva ROC.

Duración del módulo: 10 horas

Módulo 9 Visualización de Máquinas Inteligentes

Utilizar herramientas visuales de maquinas inteligentes para desarrollar modelos de datos mas exactos y/o precisos, que cumplan con los requerimientos de la necesidad del problema a resolver

Temario

1.Herramienta Yellowbrick.
2. Hyperparameters Optimization
3. Data Wrapper
4. export_graphviz from sklearn.tree
5. Decision Boundaries
6. ANN Visualiser
7. Variational Autoencoders (VAE)

Duración del módulo: 10 horas

Módulo 10 Plataformas y Máquinas Inteligentes en Big Data

Desarrollar modelos inteligentes supervisados de grandes volúmenes de datos utilizando pySpark de Python; seleccionar el modelo adecuado y analizar la exactitud, precisión del modelo, que cumplan lo mejor posible con los requerimientos de la tarea requerida.

Temario

1. Introducción
2. Estructura de datos en pySpark
3. Modelación Inteligente de datos en pySpark.

Duración del módulo: 10 horas

Módulo 11 Analítica de Texto

Desarrollar modelos inteligentes supervisado de datos de texto utilizando la plataforma máquinas inteligentes en Python; seleccionar el modelo adecuado y analizar la exactitud, precisión del modelo, que cumplan lo mejor posible con los requerimientos de la tarea requerida

Temario

1. Introducción herramientas NLTK, Spacy, TextBlob, PyTorch-NLP, Textacy.
2. Introducción: aplicaciones en detección de plagio y detección de autoría.
3. Análisis de frecuencias, personas, eventos,
4. Conceptos de Transformadores
5. Modelo inteligentes estadísticos, lingüísticos y profundos.
6. Visualización de Texto
7. Visualización del Corpus t-SNE: use stochastic neighborg embedding to project documents
8. Visualizar la dispersión de palabras clave en el Corpus.
9. Visualiza de documentos similares (UMAP)

Duración del módulo: 10 horas

Módulo 12 Analítica de Redes Sociales

Desarrollar modelos Redes de datos utilizando la NetworkX en Python; analizar la robustez de las redes, encontrar las personas que mas se comunican, así como los líderes de las conversaciones en una red social.

Temario

1. Introducción: Aplicaciones.
2. Definición de nodes, vertices, and atributos.
3. Tipos de redes: direccional, bidireccional, pesada, bipartita.
4. Representación y manipulación de datos utilizando NetworkX.
5. Métricas (distancia, alcance (reachability) y redundancia para explorar lo robusto de redes a ataques intencionales o al quitar nodos o vértices..
6. Centralidad (Grado, ?Closeness?, and ?Betweenness?, ?Page Rank?).

Duración del módulo: 10 horas