Data Science

TechGuide - Alura, FIAP e PM3

Data Science

Nivel 1

- [ ] **Ciencia de Datos - Fundamentos**:
    - La Ciencia de Datos es el acto de utilizar algoritmos y sistemas para extraer, organizar y analizar datos a partir de diversas fuentes, a fin de detectar patrones y tomar decisiones de negocios. Las áreas de aplicación son infinitas, como en negocios, biología, medicina, ingeniería, etc.
    - Conocer el concepto de Explotación de Datos
    - Conocer las principales funciones, como 'describe', 'info', etc
    - Entender el papel de las visualizaciones como histogramas y boxplots
    - Saber qué son las variables categóricas nominales y ordinales
    - Explorar los datos en Python con las bibliotecas Pandas, Matplotlib, Seaborn, etc
- [ ] **Feature Engineering**:
    - Generar nuevas variables a partir de los datos disponibles
    - Transformar datos brutos en características
    - Resaltar los problemas de los modelos predictivos
    - Mejorar la precisión del modelo para nuevos datos
    - Seleccionar y crear Features utilizando las bibliotecas Pandas y Scikit-Learn

- [ ] **Extracción y tratamiento de datos**:
  - Obtener los datos que se analizarán
  - Tratar los datos obtenidos, transformándolos, alterando su estructura y valores a fin de dejar la base de datos más coherente y garantizar que los datos que serán trabajados estén en las mejores condiciones para ser analizados
- [ ] **Python para Ciencia de Datos - Fundamentos**:
  - Python es un lenguaje de programación interpretado de alto nivel y que soporta múltiples paradigmas de programación, como imperativo, orientado a objetos y funcional. Es un lenguaje con tipificación dinámica y administración automática de memoria.
  - Aprender lógica de programación en Python
  - Aprenda los fundamentos del lenguaje como variables, funciones, listas, condicionales e Imports
  - Crear análisis de datos
  - Usar Matplotlib para generar gráficos
  - Usar y manipular listas para agrupar datos
  - Conocer la biblioteca NumPy
  - Conocer la biblioteca Pandas
- [ ] **Jupyter y Colab**:
  - Jupyter Notebook y Google Colaboratory son portátiles que permiten la creación de bloques de texto y bloques de código
  - Los Notebooks facilitan la elaboración de proyectos de Data Science por ser posible visualizar el resultado de la ejecución luego del trecho de código
  - Google Colaboratory le permite escribir y ejecutar códigos Python directamente en el navegador, sin ninguna o pocas configuraciones necesarias
  - Facilitan el intercambio de proyectos entre el equipo
- [ ] **R para Ciencia de Datos**:

- R es un lenguaje de programación comúnmente usado en estadística computacional y análisis de datos.
- Aprender a analizar, limpiar y ver datos
- Elaborar gráficos
- Unir bases de datos
- Extraer resúmenes estadísticos
- Aprender modelos predictivos en R

☐ **Estadísticas y Matemáticas - Fundamentos**:

- Para aprender técnicas de análisis de datos, estadísticas, modelado en la práctica, escalar el crecimiento y brindar nuevas oportunidades, ya sea en la empresa para la que trabaja o en su propio negocio.
- En esta formacion aprenderás utilizando el lenguaje Python y varias bibliotecas y herramientas específicas para estadísticas.
- Comprenderás cómo crear hipótesis y validarlas. De esta forma, podrá justificar los análisis realizados y mostrar a la empresa la importancia del papel del conocimiento en la estadística.
- Ecuaciones, funciones y límites
- Logaritmos
- Matrices, determinantes, vectores y espacio vectorial
- Derivadas e integrales
- Diferencia entre promedio, mediana y moda
- Distribución de frecuencia
- Varianza y desviación estándar
- Distribución binomial, de Poisson y normal
- Nivel e intervalo de confianza
- Técnicas de muestreo
- Regresión lineal

☐ **Visualización de Datos**:

- La visualización de datos es una expresión contemporánea de la comunicación visual que consiste en la representación visual de datos.
- Mapear datos abstractos en representaciones visuales
- Representar visualmente los datos que están presentes en nuestro mundo real

Nivel 2

- [ ] **Machine Learning - Fundamentos**:

- El Aprendizaje automático o Machine Learning es un subcampo de la Ingeniería y de la Ciencia de la Computación que evolucionó del estudio de reconocimiento de patrones y de la teoría del aprendizaje computacional en inteligencia artificial.
- Aprendizaje supervisado
- Utilizar algoritmos de clasificación
- Usar algoritmos de regresión
- Utilizar Scikit-Learn para crear modelos de machine Learning

- [ ] **Machine Learning - Aprendizaje no supervisado**:

- El clustering o análisis de agrupación de datos es el conjunto de técnicas de prospección de datos (data Mining) que tiene como objetivo hacer agrupaciones automáticas de datos según su grado de similitud.
- Conocer el análisis exploratorio
- Utilizar los métodos K-Means, DBSCAN y Mean shift para agrupar datos sin clasificación
- Evaluar la calidad de una Clusterización
- Parametrizar métodos de Clusterización a través del máximo coeficiente de silueta
- Entender las matemáticas detrás de las métricas de validación Silhouette, Davies Bouldin y Calinski Harabasz
- Conocer técnicas de reducción de dimensionalidad

- [ ] **Machine Learning - Evaluación de Modelos**:
  - El uso de métricas de evaluación nos ayuda a identificar si un modelo entrenado tendrá un buen desempeño de predicción cuando se expone a nuevos conjuntos de datos.
  - Conocer diferentes estrategias de evaluación y optimización de modelos
  - Utilizar una canalización para entrenamiento y validación
  - Métricas de evaluación de modelos de machine Learning
- [ ] **SQL - Fundamentos**:
  - Conocer los comandos más comunes de SQL
  - Usar SELECT para consultar una tabla
  - Usar INSERT para insertar datos en una tabla
  - Usar UPDATE para actualizar una tabla
  - Usar DELETE para eliminar datos de una tabla
  - Usar JOIN para conectar los datos de múltiples tablas
  - Conocer las cláusulas (FROM, ORDER BY, etc.)
- [ ] **Web Scraping**:
  - Web scraping o raspado de datos en la web es la extracción de datos de sitios web.
  - Usar BeautifulSoup y Python para recopilar datos
  - Buscar y navegar por HTML
  - Acceder al contenido y atributos de las etiquetas HTML
  - Construir conjuntos de datos con los resultados de los raspados
- [ ] **Pruebas Estadísticas**:
  - Crear intervalos de confianza para muestras
  - Comparar grupos de muestras
  - Realizar pruebas estadísticas
  - Planificar experimentos para la recopilación de datos

- Proponer modelos matemáticos para entender un problema dado
- Construir mapas de colores para ayudar a la interpretación de los datos

Nivel 3

- [ ] **Aprendizaje Profundo**:

- Deep Learning (o Aprendizaje Profundo) es una rama de Machine Learning basada en un conjunto de algoritmos que intentan modelar abstracciones de alto nivel de datos usando un grafo profundo con múltiples capas de procesamiento, compuestas de varias transformaciones lineales y no lineales.
- Construir y entrenar modelos con Keras
- Seleccionar las capas de una plantilla
- Clasificar imágenes
- Entender los conceptos de pesos y sesgos
- Redes neuronales para regresión

- [ ] **Aprendizaje por Refuerzo**:

- El Aprendizaje por Refuerzo es un área de Machine Learning que se preocupa con la forma como agentes inteligentes deben tomar medidas en un ambiente, a fin de maximizar la noción de recompensa acumulativa.
- Entender los conceptos de agente y recompensa
- Entender la diferencia entre refuerzo positivo y negativo
- Conocer el modelo Markov Decision Process
- Entender el concepto de Retorno
- Utilizar el algoritmo Q-Learning

- [ ] **Visión Computacional**:

- Visión Computacional es un campo científico interdisciplinario que se ocupa de cómo las computadoras pueden obtener conocimientos de alto nivel a partir de imágenes o videos digitales. Desde la perspectiva de la ingeniería,

busca comprender y automatizar tareas que el sistema visual humano puede hacer.

- Extraer regiones de interés de una imagen
- Normalizar y procesar los datos de las imágenes
- Construir clasificadores de reconocimiento facial
- Extraer regiones del rostro humano basado en hitos faciales
- Analizar diferentes condiciones de cada componente del rostro humano

- [ ] **Procesamiento de Lenguaje Natural**:

- Procesamiento de la lengua natural (PLN) es un subárea de la inteligencia artificial y la lingüística que estudia los problemas de la generación y la comprensión automática de las lenguas humanas naturales.
- Análisis de Sentimiento
- Crear vistas para facilitar el análisis de datos textuales
- Conocer las bibliotecas NLTK y Scikit-Learn
- Normalizar textos
- Usar TF-IDF y Ngrams para mejorar la clasificación
- Uso de SKlearn
- Utilizar Regex en PLN
- Conocer Word2Vec
- Combinar vectores de palabras para representar textos y clasificarlos

Habilidade Auxiliar: Cloud, Big Data y Sistemas

- [ ] **Big Data e Ingeniería de Datos - Fundamentos**:

- Big data es el área del conocimiento que estudia cómo tratar, analizar y obtener información a partir de conjuntos de datos demasiado grandes para ser analizados por sistemas tradicionales.
- La ingeniería de datos implica el desarrollo y organización de procesos para la recopilación, almacenamiento y transformación de datos a gran escala.

- Crear una canalización de datos
- Conocer el Apache Airflow
- Conocer el Apache Spark
- Conocer el concepto de Data Lake
- Interactuar con servidores en la nube
- Crear aplicaciones de Spark

☐ **Cloud - Fundamentos**:

- La computación en nube, o cloud computing, es la distribución de servicios informáticos a través de Internet mediante un modelo de tarificación de pago por uso. Una nube se compone de varios recursos informatizados, desde los propios ordenadores (o instancias, en terminología de nube) hasta las redes, el almacenamiento, las bases de datos y todo lo que les rodea. En otras palabras, todo lo que normalmente se necesita para montar el equivalente a una sala de servidores, o incluso un centro de datos completo, estará listo para usar, configurar y ejecutar.
- Conocer la diferencia entre IaaS, PaaS y SaaS
- Conocer los mayores proveedores de nube
- Especializarse en un proveedor específico de su preferencia

☐ **Git y GitHub - Fundamentos**:

- Git es un sistema de control de versiones distribuido gratuito y de código abierto diseñado para manejar todo, desde proyectos pequeños hasta proyectos muy grandes, con rapidez y eficiencia.
- GitHub es un servicio de hosting para el desarrollo de software y el control de versiones mediante Git.
- Crear un repositorio
- Clonar un repositorio
- Comprometerse, empujar y tirar hacia y desde el repositorio
- Revertir un commit
- Crear de ramas y Pull requests

- Manejar fusiones y conflictos

- [ ] **Linux - Fundamentos**:

- Linux es un término popularmente empleado para referirse a sistemas operativos que utilizan el Kernel - Linux. - Las distribuciones incluyen el kernel de Linux, además de software de sistema y bibliotecas.
- Conocer el sistema de directorios de Linux
- Comprimir y descomprimir archivos
- Administrar los procesos que se ejecutan en la máquina
- Conocer las variables de entorno y el PATH
- Administrar paquetes
- Realizar comunicación remota con SSH y SCP

Habilidade Auxiliar: Business

- [ ] **Gestión de Procesos de Negocio**:

- La Gestión de Procesos de Negocios (BPM) es la disciplina que utiliza varios métodos para descubrir, modelar, analizar, medir, mejorar, optimizar y automatizar procesos de negocios.

- [ ] **Business Intelligence (BI) - Fundamentos**:

- Business Intelligence es un conjunto de teorías, metodologías, procesos y tecnologías que posibilitan la transformación de los datos "crudos" en informaciones extremadamente relevantes para la toma de decisiones de una empresa.
- Conocer el proceso de ETL
- Realizar el modelado y estructuración de tablas en un almacén de datos
- Crear vistas que tengan sentido
- Conocer PowerBI

- [ ] **Storytelling con datos**:

- Utilizar algoritmos y sistemas para extraer, organizar y analizar datos de diversas fuentes con el fin de detectar patrones y tomar decisiones comerciales
- Las áreas de aplicación son infinitas, como en negocios, biología, medicina, ingeniería, etc.

- [ ] **Excel**:

- Microsoft Excel es un editor de hojas de cálculo producido por Microsoft con herramientas de cálculo y de construcción de tablas.
- Realizar las operaciones matemáticas básicas con sus operadores (suma, resta, multiplicación y división).
- Conocer las principales fórmulas, como 'MEDIA' (AVERAGE), 'ARRED' (ROUND), 'MÁXIMO' (MAX), 'MÍNIMO' (MIN), etc.
- Realizar búsquedas en columnas con la función 'PROCV'.

- [ ] **Habilidades de Comunicación**:

- Un buen nivel de comunicación facilita el logro de objetivos, resolución de problemas, además de aumentar la productividad, porque cada profesional sabrá exactamente lo que se espera de él y transmitir con claridad sus ideas.
