Curso

Fundamentos Data Science

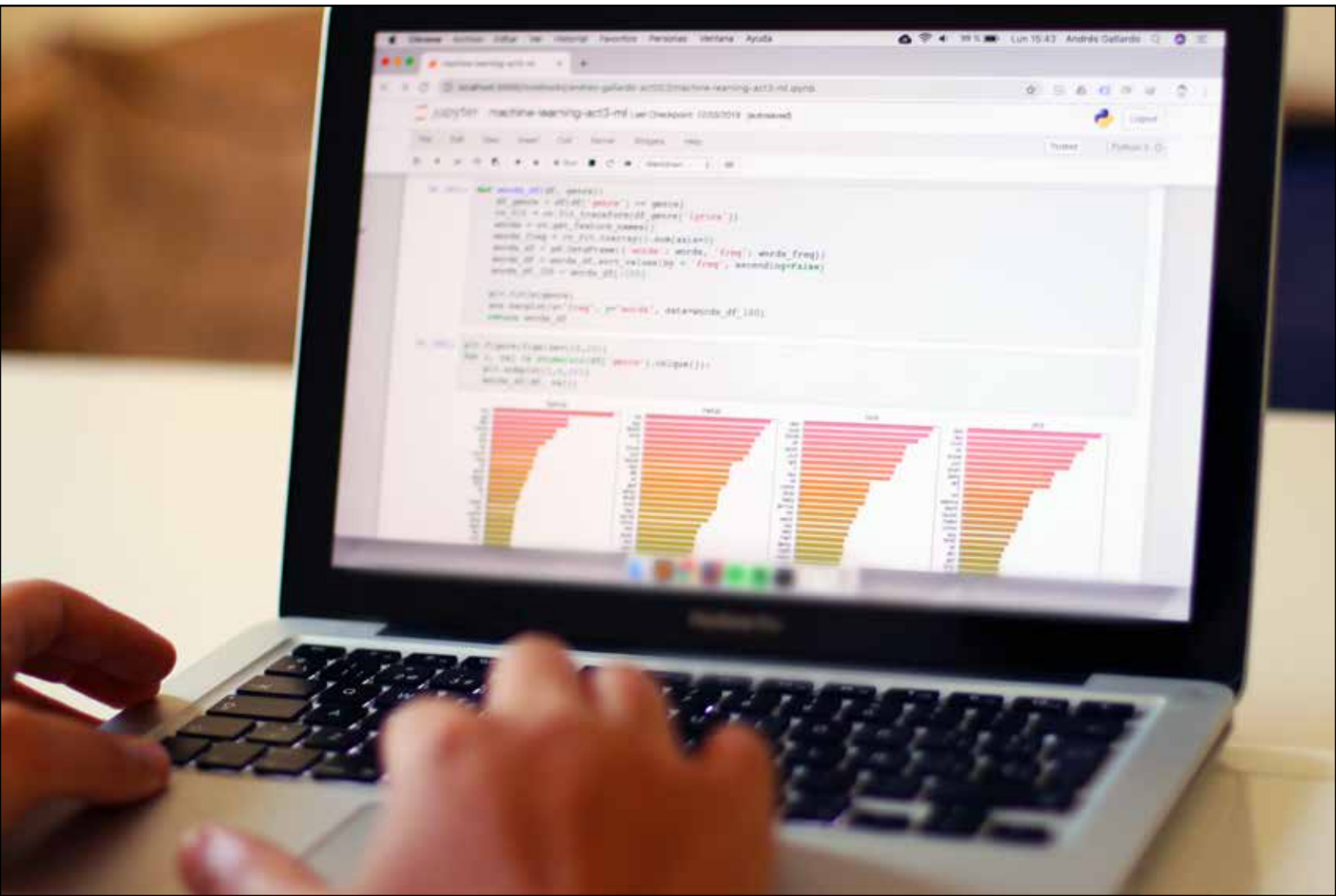
{desafío} latam_

Descripción del **Programa**

El curso de Fundamentos de Data Science te entrega los elementos fundacionales de la ciencia de datos en cuanto a habilidades de programación y modelación estadística. A lo largo del curso aprenderás a manipular datos y solicitar información mediante Python y las principales librerías asociadas al trabajo como Pandas, Numpy, Scipy, Matplotlib, StatsModels y Scikit-Learn.

Explorarás dos de las principales tradiciones analíticas que dominan el rol de los científicos de datos, la econometría y el aprendizaje de máquinas (machine learning) entregando la base teórica y las competencias necesarias para generar aproximaciones a la información disponible, acorde a los requerimientos de la industria.

Este programa es el segundo módulo de la carrera de Data Science de Desafío Latam.



Unidades y **Contenidos**

Unidad 1

Estadística Univariada y Control de Flujo

- Conocer los principales modos de trabajo con Jupyter Notebook.
- Utilizar las estructuras de datos de pd.Series y pd.DataFrame.
- Analizar datos de forma univariada con pandas.
- Utilizar control flujos para obtener medidas estadísticas.

Unidad 2

Probabilidades y Funciones

- Utilizar funciones para reutilizar código. (Principio D.R.Y)
- Convertir una fórmula matemática a una función en Python.
- Construir y utilizar funciones orientadas al análisis de datos.
- Optimizar funciones reemplazandolas por funciones vectorizadas.
- Utilizar conceptos básicos de probabilidad.
- Generar segmentaciones de un pd.DataFrame en base a indexación y selección.

Unidad 3

Variables Aleatorias y Gráficos

- Hacer uso de métodos de pandas para segmentar columnas y filas.
- Hacer uso de los métodos iterrows e iteritems para implementar loops en pandas.
- Implementar enumerate en loops.
- Conocer las convenciones y principios rectores de la visualización de gráficos.
- Conocer las principales convenciones en la visualización de resultados en histogramas, gráficos de punto y barras.
- Generar simulaciones de la distribución normal.
- Conocer las principales aplicaciones de las distribuciones.
- Calcular e interpretar puntajes z.
- Describir la Ley de los Grandes Números y Teorema del Límite Central y su importancia en la inferencia estadística.

Unidad 4

Hipótesis y Correlación

- Conocer las funcionalidades avanzadas de gráficos estáticos mediante seaborn.
- Aprender a segmentar datos y los principales criterios de estratificación.
- Conocer los principales criterios de transformación de variables.
- Aplicar funciones a columnas de datos mediante ufuncs, map-reduce-filter.
- Entender e interpretar la correlación a partir de diagramas de dispersión.
- Entender el marco inferencial frecuentista de las hipótesis.
- Conocer la distribución t de Student y su aplicación.
- Aplicar pruebas de hipótesis simples en el contexto de la inferencia.

Unidad 5

Regresión

- Reconocer la terminología asociada a la modelación estadística.
- Conocer la regresión lineal y sus fundamentos.
- Interpretar los parámetros estimados en la regresión.
- Conocer y ser capaz de interpretar estadísticos de bondad de ajuste y coeficientes.
- Reconocer los supuestos en los que la regresión tiene sustento teórico.
- Implementar un modelo de regresión con statsmodels.
- Utilizar transformaciones simples en las variables independientes.
- Implementar un modelo predictivo con scikit-learn.

Unidad 6

Clasificación

- Conocer la regresión logística y sus fundamentos.
- Conocer y ser capaz de interpretar estadísticos de bondad de ajuste y coeficientes.
- Reconocer los supuestos en que tiene sustento teórico.
- Implementar un modelo de regresión con statsmodels.
- Implementar un modelo predictivo con scikit-learn.
- Conocer los conceptos de validación cruzada y medidas de desempeño.

Unidad 7

Dimensionalidad y Agrupación

- Entender el problema de la "maldición de la dimensionalidad" y sus implicancias para el modelo.
- Conocer la aproximación psicométrica del
- Principal Component Analysis y el Análisis Factorial.
- Implementar algoritmos de reducción de dimensiones (Principal Components Analysis) y de reconocimiento de estructuras latentes (Análisis Factorial) con scikit-learn.
- Utilizar técnicas para identificar patrones de datos perdidos.
- Implementar algoritmos de agrupación (k-Means).

Unidad 8

Modelos Generalizados

- Conocer los componentes del marco analítico de los Modelos Lineales
- Generalizados (Componentes estocásticos, sistemáticos y funciones de enlace).
- Conocer el método de estimación por Máxima Verosimilitud con el que se estiman los Modelos Lineales Generalizados.
- Identificar la correcta implementación de los modelos en base a la naturaleza del problema.
- Implementar modelos mediante la librería statsmodels acorde a la naturaleza del problema.
- Interpretar las estimaciones de manera correcta tomando en cuenta las funciones de enlace asociadas a cada modelo.

Duración

- **8 semanas**
- **Sesión online:**
  48 horas (8 sesiones de 6 horas cada una)
- **Sesión presencial:**
  48 horas (16 sesiones de 3 horas cada una)

Requerimientos

Características de tu notebook*

- Sistema Operativo: Windows, Linux o Mac
- Procesador Intel Core i3, 8GB RAM, 128 Disco SSD

Plataformas y Software

- Empieza (https://empieza.desafiolatam.com)
- Jupyter Notebook
- iPython Kernel
- Anaconda

** El notebook es por cuenta de todos los participantes: docente, ayudante y alumnos.*
***Programas open source, por lo que el estudiante no necesita incurrir en gastos de licencias.*

Curso
Fundamentos Data Science

{desafío} latam_

www.desafiolatam.com

/DesafioLatam /DesafioLatam /DesafioLatam /DesafioLatam