

## PROGRAMA ACADÉMICO

<b>Módulo</b>	Fundamentos Data Science
<b>Descripción</b>	<p>El módulo Fundamentos de Data Science entrega los elementos fundacionales de la ciencia de datos en cuanto a habilidades de programación y modelación estadística. A lo largo del curso el alumno aprenderá a manipular datos y solicitar información mediante Python y las principales librerías asociadas al trabajo como pandas, numpy, scipy, matplotlib, statsmodels y scikit-learn.</p> <p>También serán expuestos a las dos 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.</p>
<b>Competencias Generales</b>	<ul style="list-style-type: none"><li>• Conocer los elementos fundacionales del análisis estadístico y de la programación orientada a la estadística.</li><li>• Aplicar métodos estadísticos para extraer información descriptiva y generar inferencias con datos limitados.</li><li>• Adquirir capacidades analíticas básicas desde la econometría y machine learning.</li><li>• Utilizar Jupyter Notebook para generar reportes y visualizaciones sobre datasets.</li></ul>

## ALCANCE

Unidad	Descripción	Competencias
Estadística Univariada y Control de Flujo	Generar mediciones descriptivas de una base de datos (media, moda, mediana, desviación estándar) e implementar control de flujo para estructuras de datos de Pandas.	<ul style="list-style-type: none"><li>• Conocer los principales modos de trabajo con Jupyter Notebook.</li><li>• Utilizar las estructuras de datos de pd.Series y pd.DataFrame.</li><li>• Analizar datos de forma univariada con pandas.</li><li>• Utilizar control flujos para obtener medidas estadísticas.</li></ul>
Probabilidades y Funciones	Generar mediciones de probabilidad para analizar la frecuencia de ocurrencia de un fenómeno específico mediante Python.  Implementar funciones en el contexto de las librerías pandas y numpy.	<ul style="list-style-type: none"><li>• Utilizar funciones para reutilizar código. (Principio D.R.Y)</li><li>• Convertir una fórmula matemática a una función en Python.</li><li>• Construir y utilizar funciones orientadas al análisis de datos.</li><li>• Optimizar funciones reemplazandolas por funciones vectorizadas.</li><li>• Utilizar conceptos básicos de probabilidad.</li><li>• Generar segmentaciones de un pd.DataFrame en base a indexación y selección.</li></ul>
Variables Aleatorias y Gráficos	Bases teóricas para entender cómo podemos aproximarnos a fenómenos empíricos mediante leyes matemáticas, así como aprender a graficar utilizando la librería matplotlib.	<ul style="list-style-type: none"><li>• Hacer uso de métodos de pandas para segmentar columnas y filas.</li><li>• Hacer uso de los métodos iterrows e iteritems para implementar loops en pandas.</li><li>• Implementar enumerate en loops.</li><li>• Conocer las convenciones y principios rectores de la visualización de gráficos.</li><li>• Conocer las principales convenciones en la visualización de resultados en histogramas, gráficos de punto y</li></ul>

		<p>barras.</p> <ul style="list-style-type: none"> <li>• Generar simulaciones de la distribución normal.</li> <li>• Conocer las principales aplicaciones de las distribuciones.</li> <li>• Calcular e interpretar puntajes z.</li> <li>• Describir la Ley de los Grandes Números y Teorema del Límite Central y su importancia en la inferencia estadística.</li> </ul>
Hipótesis y Correlación	Formular y poner a prueba hipótesis desde el marco analítico frecuentista, así como analizar la relaciones existentes entre dos variables mediante los diagramas de dispersión y la correlación.	<ul style="list-style-type: none"> <li>• Conocer las funcionalidades avanzadas de gráficos estáticos mediante seaborn.</li> <li>• Aprender a segmentar datos y los principales criterios de estratificación.</li> <li>• Conocer los principales criterios de transformación de variables.</li> <li>• Aplicar funciones a columnas de datos mediante ufuncs, map-reduce-filter.</li> <li>• Entender e interpretar la correlación a partir de diagramas de dispersión.</li> <li>• Entender el marco inferencial frecuentista de las hipótesis.</li> <li>• Conocer la distribución t de Student y su aplicación.</li> <li>• Aplicar pruebas de hipótesis simples en el contexto de la inferencia.</li> </ul>
Regresión	Conocimientos teóricos y prácticos sobre la regresión desde la econometría y el machine learning, comprendiendo la terminología y los principales pasos en el flujo de trabajo de	<ul style="list-style-type: none"> <li>• Reconocer la terminología asociada a la modelación estadística.</li> <li>• Conocer la regresión lineal y sus fundamentos.</li> <li>• Interpretar los parámetros estimados en la regresión.</li> <li>• Conocer y ser capaz de interpretar estadísticos de bondad de ajuste y</li> </ul>

	machine learning.	<p>coeficientes.</p> <ul style="list-style-type: none"> <li>• Reconocer los supuestos en los que la regresión tiene sustento teórico.</li> <li>• Implementar un modelo de regresión con statsmodels.</li> <li>• Utilizar transformaciones simples en las variables independientes.</li> <li>• Implementar un modelo predictivo con scikit-learn.</li> </ul>
Clasificación	Conocer los modelos de clasificación (desde la econometría y el machine learning), comprender en qué situaciones se pueden implementar sus soluciones y conocer los métodos de validación cruzada.	<ul style="list-style-type: none"> <li>• Conocer la regresión logística y sus fundamentos.</li> <li>• Conocer y ser capaz de interpretar estadísticos de bondad de ajuste y coeficientes.</li> <li>• Reconocer los supuestos en que tiene sustento teórico.</li> <li>• Implementar un modelo de regresión con statsmodels.</li> <li>• Implementar un modelo predictivo con scikit-learn.</li> <li>• Conocer los conceptos de validación cruzada y medidas de desempeño.</li> </ul>
Dimensionalidad y Agrupación	introducción en la problemática asociada cuando se tienen más atributos que observaciones, y entregar soluciones orientadas a la resolución y disminución de dimensiones.	<ul style="list-style-type: none"> <li>• Entender el problema de la "maldición de la dimensionalidad" y sus implicancias para el modelo.</li> <li>• Conocer la aproximación psicométrica del Principal Component Analysis y el Análisis Factorial.</li> <li>• Implementar algoritmos de reducción de dimensiones (Principal Components Analysis) y de reconocimiento de estructuras latentes (Análisis Factorial) con scikit-learn.</li> <li>• Utilizar técnicas para identificar patrones de datos perdidos.</li> </ul>

		<ul style="list-style-type: none"> <li>• Implementar algoritmos de agrupación (k-Means).</li> </ul>
Modelos Generalizados	Marco analítico de los modelos lineales generalizados, mediante el cual pueden aproximarse a distintos problemas cuando el vector objetivo no sigue una distribución normal.	<ul style="list-style-type: none"> <li>• Conocer los componentes del marco analítico de los Modelos Lineales Generalizados (Componentes estocásticos, sistemáticos y funciones de enlace).</li> <li>• Conocer el método de estimación por Máxima Verosimilitud con el que se estiman los Modelos Lineales Generalizados.</li> <li>• Identificar la correcta implementación de los modelos en base a la naturaleza del problema.</li> <li>• Implementar modelos mediante la librería statsmodels acorde a la naturaleza del problema.</li> <li>• Interpretar las estimaciones de manera correcta tomando en cuenta las funciones de enlace asociadas a cada modelo.</li> </ul>

## REQUERIMIENTOS

<b>Recursos de Información**</b>	<b>Características notebook*</b> <ul style="list-style-type: none"><li>• Sistema Operativo: Linux o Mac</li><li>• Procesador: Intel Core i3</li><li>• RAM: 8GB, 128 Disco SSD</li></ul> <b>Plataformas y Software</b> <ul style="list-style-type: none"><li>• LMS</li><li>• Jupyter Notebook</li><li>• iPython Kernel</li><li>• Anaconda</li></ul>
----------------------------------	--

*\* 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.*