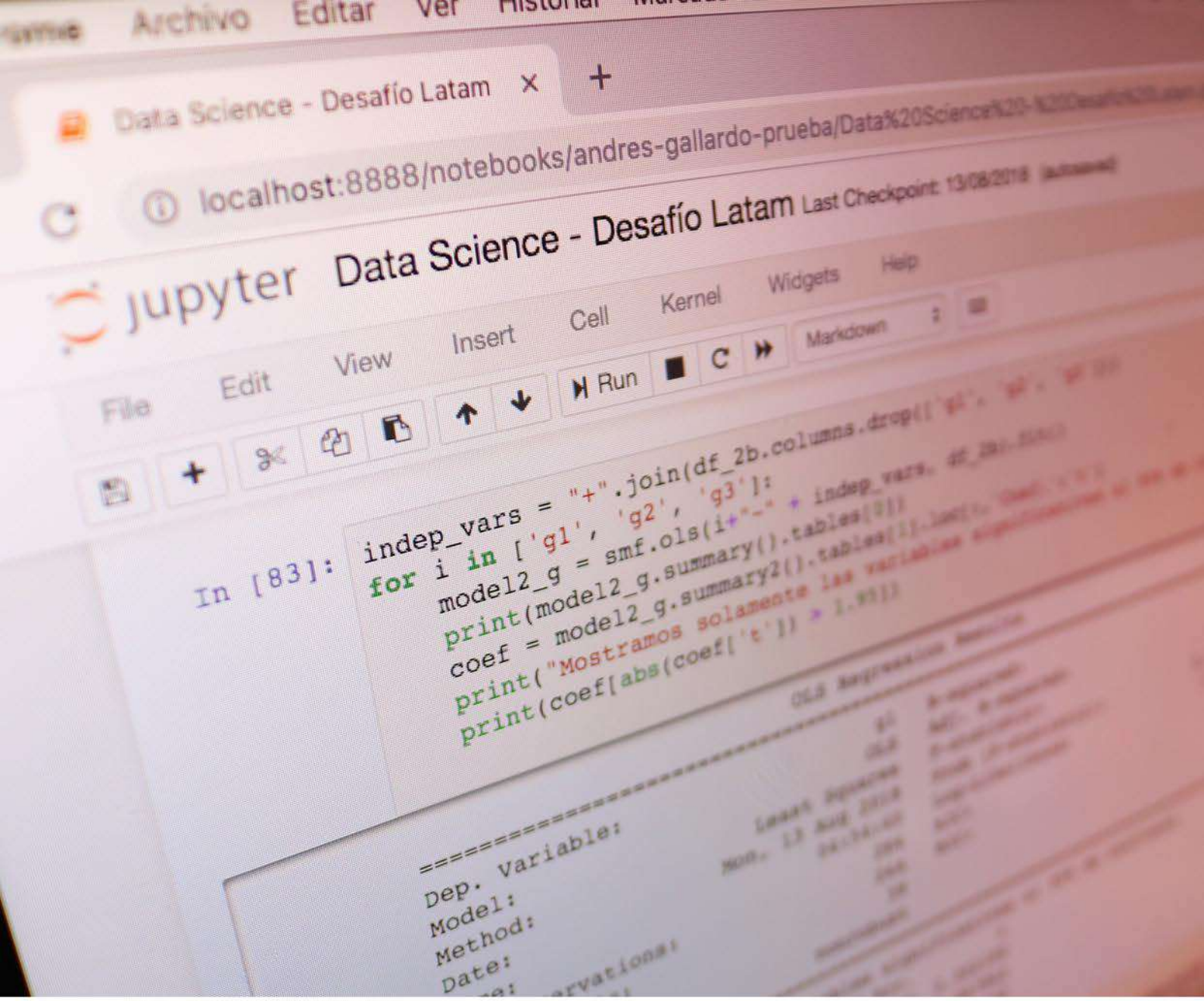


PROGRAMA FORMATIVO

DATA SCIENCE

Octubre 2018

ACADEMIA DESAFIO LATAM

Plan Formativo

Nombre	Data Science
Duración	31 semanas • Sesión online **168 horas** • Sesión presencial **186 horas**
Descripción	El plan formativo del Data Scioence consta de 6 módulos: Introducción a la Programación con Python, Fundamentos de Data Science, Machine Learning, SQL para Data Science, Big Data y Proyecto Data Science. Los contenidos se dividen en dos sesiones: • **Sesión online:** El contenido teórico se encuentra disponible en la plataforma Empieza de la academia Desafío Latam, este contenido al estar disponible de manera online entrega flexibilidad a los alumnos para que puedan avanzar a su propio ritmo. Se recomiendan 6 horas de autoestudio por parte del alumno por semana de contenidos. Además del contenido teórico, la plataforma contiene: Pautas de evaluación, control de asistencia, actividades prácticas (desafíos) y estadísticas. • **Sesión presencial:** Las clases prácticas consisten en dos sesiones por semana con una duración de 3 horas cada una y se centran en actividades prácticas en las cuales los alumnos desarrollan desafíos reales con el acompañamiento del docente, ayudante y compañeros. Incentivando la innovación, trabajo en equipo, participación activa y rápido aprendizaje.
Perfil de Egreso	El egresado de Data Science cubre la creciente necesidad de

	la sociedad de comprender y analizar grandes cantidades de datos que se generan, ya que posee las herramientas teóricas y prácticas para implementar modelos descriptivos y predictivos acorde a las diversas áreas de la industria de acuerdo a la naturaleza de los datos. Podrá manejar y extraer la información, facilitando la lectura del código para apoyar en la toma de decisiones que mejoren y optimicen la calidad de vida.
Requisitos de Ingreso	El alumno requiere de conocimientos previos en las áreas de matemática y estadística. Cada uno de ellos responde a la naturaleza híbrida de la ciencia de datos: **Matemática y estadística de formación general:** • Plano cartesiano y ecuación de la recta. • Manejo de funciones lineales y cuadráticas. • Manejo de logaritmos y funciones. • Estadística univariada (concepto de media, moda y varianza). • Probabilidad básica (fracciones, porcentajes, frecuencias relativas y absolutas, concepto de eventos dependientes e independientes). Para la validación de los conocimientos previos antes descritos, se realizará una prueba de selección múltiple.

Módulo 1: Introducción a la programación con Python

Número de horas	• Sesión online: **24 horas** (4 sesiones de 6 horas cada una) • Sesión presencial: **24 horas** (8 sesiones de 3 horas cada una)
Descripción	El módulo introductorio de programación con `Python` entrega las herramientas y conocimientos básicos para construir scripts que puedan leer datos desde archivos y otras fuentes de información, limpiar datos de acuerdo a las necesidades del negocio y generar archivos finales, fáciles de procesar, habilitando en el manejo de herramientas y flujos de trabajo necesarios para posteriormente realizar análisis estadísticos sobre los datos trabajados.
Competencias	• Instalar un entorno de trabajo para `Python3`. • Conocer las reglas sintácticas de `Python`. • Conocer el flujo de trabajo para procesar datos desde un archivos y crear un nuevo archivo con los datos modificados. • Construir scripts reutilizables en `Python` para procesar datos.

Alcance

UNIDAD	Nº SESIONES	ALCANCE
1. Introducción a la programación	Online: 1 Presencial: 2	• Instalar herramientas: Editor, Python, Python interactivo, Pip. • Ejecutar Python desde el terminal. • Ejecutar Python desde el editor de texto. • Realizar diagramas de flujo y pseudocódigo. • Construir aplicaciones tipo calculadora. • Manejar flujos y operadores lógicos.
2. Ciclos y métodos	Online: 1 Presencial: 2	• Leer y transcribir diagramas de flujo con interacciones a código Python. • Analizar e implementar diagramas de flujo con repeticiones. • Identificar componentes de un flujo **`for`**. • Validar entradas de un iterador. • Conocer el concepto de complejidad algorítmica y sus implicaciones para el desarrollo de flujos. • Conocer la notación Big-O para el cálculo de complejidades en función a la cantidad de ciclos. • Identificar los elementos que componen una función. • Conocer el scope de una función. • Identificar el alcance de variables globales y locales. • Seguir la orden de ejecución de una función mediante **`pdb`**. • Entender el retorno implícito **`return None`** en Python. • Implementar retornos explícitos en la función. • Conocer el principio *Don't Repeat Yourself*

		y su relevancia para la implementación de código. • Identificar los argumentos por defecto y opcionales. • Conocer **args** y **kwargs**. • Manejar excepciones. • Conocer la diferencia entre error y excepción. • Conocer los antipatrones utilizados en las excepciones. • Implementar excepciones mediante **try** y **except**. • Realizar debugging. • Dividir un proyecto en varios archivos. • Instalar componentes vía **PIP**.
3. Estructuras de datos	Online: 1 Presencial: 2	• Diferenciar las estructuras de datos de datos básicos de Python: **list**, **tuple** y **set**. • Manejar listas. • Resolver problemas típicos de listas: reducciones, transformaciones, filtros y selecciones y lectura de datos desde archivos. • Realizar persistencia.
4. API	Online: 1 Presencial: 2	• Guardar datos en archivos. • Entender el objetivo de una API. • Conocer los principales componentes de una API. • Conocer las herramientas que utilicen API para la interacción (Postman). • Conocer la lógica de consumo de API. • Utilizar Postman para realizar requests a una API. • Conocer los endpoints. • Conocer y entender los verbos REST. • Conocer el formato de archivos JSON. • Utilizar Python para realizar un request a

		una API. • Consumir los datos desde Python. • Guardar los requests de una API en un JSON y procesarlos dentro de Python. • Conocer y comprender la importancia del protocolo HTTPS. • Utilizar Python para realizar un request mediante HTTPS.

Estrategia Evaluativa

La estrategia de evaluación considera:

Quizzes: Preguntas de alternativa múltiple durante las sesiones online.

Desafíos: Actividades prácticas que se realizan en las sesiones presenciales. Estas son instancias de implementación de los conocimientos adquiridos en las sesiones online, estos permiten identificar el avance de los alumnos respecto a los aprendizajes esperados del módulo, adquiriendo las competencias señaladas.

Módulo 2: Fundamentos de Data Science

Número de horas	• Sesión online: **48 horas** (8 sesiones de 6 horas cada una) • Sesión presencial: **48 horas** (16 sesiones de 3 horas cada una)
Descripción	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`. 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.
Competencias	Las competencias generales del módulo están orientadas a generar un entendimiento sólido en los elementos fundacionales de la ciencia de datos: • Conocer los elementos fundacionales del análisis estadístico y de la programación orientada a la estadística. • Aplicar métodos estadísticos para extraer información descriptiva y generar inferencias con datos limitados. • Adquirir capacidades analíticas básicas desde la econometría y machine learning. • Utilizar `Jupyter Notebook` para generar reportes y visualizaciones sobre datasets.

Alcance

UNIDAD	Nº SESIONES	ALCANCE
1. Estadística Univariada y Control de Flujo. 1.1. Introducción al ambiente de trabajo. 1.2. Control de Flujo para la estadística univariada.	Online: 1 Presencial: 2	• 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.
2. Probabilidades y Funciones. 2.1. Cálculo de probabilidades. 2.2. Implementar funciones para el cálculo de probabilidades.	Online: 1 Presencial: 2	• 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.
3. Variables Aleatorias y Gráficos. 3.1. Distribución normal. 3.2. Distribución discreta.	Online: 1 Presencial: 2	• 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.
4. Hipótesis y Correlación. 4.1. Refactorizar gráficos y relaciones bivariadas. 4.2. Generar pruebas de hipótesis.	Online: 1 Presencial: 2	• 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.
5. Regresión. 5.1. Regresión desde la econometría. 5.2. Regresión desde machine learning.	Online: 1 Presencial: 2	• 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`.
6. Clasificación. 6.1. Clasificación desde la econometría. 6.2. Clasificación desde machine learning.	Online: 1 Presencial: 2	• 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.
7. Dimensionalidad y Agrupación. 7.1. Dimensionalidad y la maldición de la dimensionalidad. 7.2. Agrupación y métodos no-supervisados.	Online: 1 Presencial: 2	• 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).

8. Modelos Generalizados. 8.1 Modelos Lineales Generalizados.	Online: 1 Presencial: 2	• 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.

Estrategia Evaluativa

La estrategia de evaluación considera:

Quizzes: Preguntas de alternativa múltiple durante las sesiones online.

Desafíos: Son actividades prácticas que se realizan en las sesiones presenciales. Estas son instancias de implementación de los conocimientos adquiridos en las sesiones online, estos permiten identificar el avance de los alumnos respecto a los aprendizajes esperados del módulo, adquiriendo las competencias señaladas.

Prueba: En la última semana del módulo, los alumnos tendrán que desarrollar una prueba práctica que aborda todo lo aprendido y aplicado durante el módulo.

Módulo 3: Machine Learning

Número de horas	• Sesión online: **48 horas** (8 sesiones de 6 horas cada una) • Sesión presencial: **48 horas** (16 sesiones de 3 horas cada una)
Descripción	El módulo Machine Learning profundiza en el aprendizaje de máquinas, entregando herramientas teóricas y prácticas al cómo permitimos que las computadoras aprendan a generalizar comportamientos en base a la información suministrada. Mediante la exposición a modelos, algoritmos de predicción y patrones, se estará preparado para la selección y preparación de un flujo de trabajo, implementando modelos predictivos para diversos casos.
Competencias	Las competencias generales para este módulo están enfocadas a la implementación eficaz de flujos de trabajo en aprendizaje automatizado: • Adquirir conocimiento y buenas prácticas de modelos predictivos. • Conocer los principales algoritmos y patrones de análisis. • Aplicar flujos de trabajo para implementar modelos predictivos. • Generar propuestas de análisis dependiendo del tipo de problema. • Predecir e inferir patrones en el comportamiento de distintos tipos de datos utilizando diversos enfoques como ensambles y redes neuronales.

Alcance

UNIDAD	Nº SESIONES	ALCANCE
1. Regularización y Expansiones Basales 1.1. Métodos de regularización: Ridge, Lasso y Elastic Net. 1.2. Manejo de no-linealidades: Modelos Generalizados Aditivos (GAM) y Regresión Multivariada con Splines adaptativos (MARS)	Online: 1 Presencial: 2	• Conocer la mecánica de operación en Ridge y Lasso. • Utilizar los métodos de regularización para resolver problemas de dimensionalidad y mejorar el desempeño predictivo. • Implementar los métodos de regularización Ridge y Lasso con `scikit-learn`. • Visualizar el comportamiento de las variables en los métodos Ridge y Lasso. • Conocer Elastic Net como un método híbrido entre las penalizaciones de los métodos de regularización Ridge y Lasso. • Importar los módulos `pyGAM` y `pyearth`. • Conocer los principales usos de GAM y MARS para flexibilizar parámetros. • Implementar los modelos GAM y MARS.

2. Modelos de Clasificación 2.1. Bayes Ingenuo 2.2. Modelos Discriminantes Lineales (LDA) y Cuadráticos (QDA) 2.3. Máquinas y Regresión de Soporte Vectoriales, Kernelización, 2.4. Algoritmo Esperanza-Maximización: Mezcla de Gaussianas, Maximización de Entropía, Análisis de Clases Latentes.	Online: 2 Presencial: 4	• Identificar los componentes del Teorema de Bayes. • Reconocer el problema de probabilidad inversa y su solución con Bayes Ingenuo • Generar funciones propias para implementar Bayes Ingenuo. • Implementar Bayes Ingenuo con `scikit-learn`. • Identificar y resolver el problema multiclases que LDA y QDA solucionan. • Entender las virtudes de LDA y QDA por sobre la regresión logística. • Reconocer los componentes paramétricos. • Implementar modelos con `scikit-learn`. • Maximizar margen de separación. • Utilizar variables Hinge y Slack en el problema de separación. • Conocer las ventajas en memoria y cantidad de parámetros estimados de SVM y SVR. • Implementar el truco de kernelización para casos N-dimensionales. • Exponer el algoritmo de Esperanza-Maximización como técnica iterativa para encontrar parámetros de Máxima Verosimilitud o Máximo Posterior.. • Aplicar Esperanza-Maximización en problemas de clases latentes e imputación múltiple. • Revisitar la regresión multinomial desde la maximización de entropía.
3. Ensambles 3.1. Árboles de Decisión.	Online: 2 Presencial: 4	• Entender el proceso de búsqueda binaria en los árboles de decisión (clasificación y regresión). • Reconocer los conceptos de pureza,

3.2. Random Forests. 3.3. Bagging. 3.4. Boosting y Gradient Boosting.		ganancia y entropía. • Implementar árboles de clasificación con `scikit-learn`. • Identificar los problemas que pueden ocurrir en la clasificación. • Reconocer la estrategia de remuestreo en el conjunto de entrenamiento y diferenciarla de la validación cruzada. • Entender los Random Forest como un híbrido entre bagging y árboles de decisión. • Implementar modelos con `scikit-learn`. • Conocer la técnica como el promedio de una serie de clasificadores débiles. • Implementar boosting mediante modelos de regresión y árboles. • Implementar algoritmos `Adaboost` y `XGBoost`.
4. Redes Neuronales. 4.1. Métodos de optimización 4.2. Perceptron y Tensores 4.3. `Tensorflow` y `Keras` 4.4. Multilayer 4.5. Convultional 4.6. Rejoinder	Online: 3 Presencial: 6	• Conocer el problema de la optimización y su terminología. • Reconocer las principales variantes del método descenso de gradiente (Normal, Batch, Mini, Stochastic) • Implementar y analizar las ventajas de cada método con `scikit-learn`. • Reconocer los orígenes y terminología asociada a las redes neuronales. • Identificar los componentes básicos de una red neuronal artificial: input, hidden, output. • Implementar un perceptron con `Python` nativo. • Entender el uso de los ambientes virtualizados (`virtualenvs`) para evitar conflicto entre librerías. • Conocer `Tensorflow` y sus principales

		elementos y funcionalidades. • Utilizar Keras como una interfaz de bajo nivel entre el usuario y `Tensorflow`. • Implementar una red neuronal básica con `Tensorflow`. • Conocer e implementar las principales funciones de activación neuronal. • Identificar las estrategias feed forward y backpropagation. • Implementar entrenamiento de redes neuronales con distintos algoritmos: `Adagrad`, `Adadelta`, `Adam`, `Nesterov` • Conocer e implementar redes neuronales multilayer y convultional.

Estrategia Evaluativa

La estrategia de evaluación considera:

Quizzes: Preguntas de alternativa múltiple durante las sesiones online.

Desafíos: Son actividades prácticas que se realizan en las sesiones presenciales. Estas son instancias de implementación de los conocimientos adquiridos en las sesiones online, estos permiten identificar el avance de los alumnos respecto a los aprendizajes esperados del módulo, adquiriendo las competencias señaladas.

Prueba: En la última semana del módulo, los alumnos tendrán que desarrollar una prueba práctica que aborda todo lo aprendido y aplicado durante el módulo.

Módulo 4: SQL para Data Science

Número de horas	• Sesión online: **18 horas** (3 sesiones de 6 horas cada una) • Sesión presencial: **18 horas** (6 sesiones de 3 horas cada una)
Descripción	En el módulo SQL para Data Science el alumno estudia el rol que juegan las bases de datos relacionales dentro del big data y sus límites. Desde el enfoque de datos relacionales se estudia el proceso de modelamiento y construcción de una base de datos relacional y la captura de datos para realizar análisis y exportación de datos para su posterior procesamiento con las herramientas aprendidas en los módulos Fundamentos de Data Science y Machine Learning.
Competencias	• Conocer el alcance de los motores de bases de datos relacionales dentro del Big Data. • Configurar un motor de base de datos `postgreSQL`. • Instalar y utilizar herramientas para facilitar la gestión de `PostgreSQL`. • Utilizar `SQL` para obtener datos de tablas y realizar operaciones matemáticas, agrupamiento y algebraicas para obtener información de datos estructurados. • Utilizar `SQL` para insertar y modificar datos de una o más tablas de una base de datos. • Acelerar las búsquedas con estrategias de indexación de contenido. • Aprender a crear bases de datos y tablas. • Realizar consultas utilizando `SQL`.

Alcance

UNIDAD	Nº SESIONES	ALCANCE
1. Modelamiento y gestión de base de datos 1.1. Modelamiento de base de datos. 1.2. Gestión de base de datos en `postgreSQL`	Online: 1 Presencial: 2	• Reconocer entidades y atributos. • Generar modelos lógicos y físicos diagramando relaciones. • Instalar y configurar `postgreSQL`. • Instalar y utilizar `pgAdmin4` y sus herramientas para facilitar la gestión de `postgreSQL`. • Construir bases de datos, tablas, índices y relaciones. • Cargar datos en formato `SQL` y dump. • Exportar datos a formato `csv`.
2. Seguridad e Integridad 2.1. Seguridad. 2.2. Integridad.	Online: 1 Presencial: 2	• Conocer reglas de acceso a una base de datos. • Aplicar constraints y técnicas de indexación para asegurar la integridad de entidad e integridad referencial.
3. Conexión de base de datos 3.1. Conexión de bases de datos con `postgreSQL`	Online: 1 Presencial: 2	• Cargar directamente a `Python` entre bases de datos. • Implementar las librerías `psycopg2` y `sqlalchemy` para la carga de bases de datos.

Módulo 5: Big Data

Número de horas	• Sesión online: **30 horas** (5 sesiones de 6 horas cada una) • Sesión presencial: **30 horas** (10 sesiones de 3 horas cada una)
Descripción	El módulo Big Data permite al alumno clasificar los problemas de big data según sus características, dimensionar según su volumen. Además podrá escoger las estrategias y herramientas adecuadas para procesar los datos dependiendo de su volumen, utilizando herramientas como `Hadoop`, `Apache Spark` y los servicios distribuidos en las nube de Amazon, para analizar grandes flujos de datos sin las limitaciones de un ambiente centralizado.
Competencias	• Determinar la necesidad de ocupar herramientas de Big Data dada las dimensiones de un problema. • Crear un flujo de `MapReduce` para resolver un problema de Big Data. • Utilizar `Hadoop` en un ambiente centralizado para generar soluciones prototipos a un problema de Big Data. • Crear instancias en `Amazon Web Services` para generar un cluster y aplicar la soluciones creadas previamente en el entorno local para trabajar con grandes cantidades de datos. • Conocer las ventajas de `Apache Spark` por sobre `Hadoop`. • Crear RDDs para resolver problemas de `MapReduce`. • Levantar un cluster de `Apache Spark` en `Amazon Web Services`.

Alcance

UNIDAD	Nº SESIONES	ALCANCE

1. Introducción a Big Data.	Online: 1 Presencial: 2	• Conocer los conceptos de complejidad algorítmica, serialización, tipos y tamaños de datos. • Reconocer las características de Big Data. • Identificar las principales diferencias de Big Data con las técnicas de los módulos Fundamentos Data Science, Machine Learning y SQL para Data Science. • Identificar los tipos de soluciones de Big Data.
2. Preparación del ambiente de trabajo	Online: 1 Presencial: 2	• Conocer los servicios AWS y CGP para el Big Data. • Crear ambientes de trabajo de manera local. • Crear ambientes de trabajo mediante AWS de manera local. • Crear y ejecutar scripts para implementar trabajos con datasets reales. • Conocer el procesamiento masivo y distribuido de datos. • Identificar las principales soluciones de procesamiento masivo y distribuido de datos.
3. Introducción a Hadoop	Online: 1 Presencial: 2	• Conocer el Ecosistema Hadoop y sus principales componentes. • Identificar los tipos de memoria y su jerarquía. • Conocer la arquitectura de Hadoop, YARN y HDFS. • Implementar mappers y reducers siguiendo el paradigma MapReduce. • Identificar e implementar las principales operaciones de procesamiento de datos en Hadoop. • Reconocer e implementar las principales

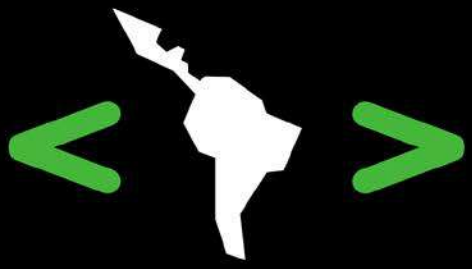
		operaciones básicas de optimización.
4. Hive	Online: 1 Presencial: 2	• Conocer la arquitectura de Hive y sus principales componentes. • Identificar en qué situaciones se puede implementar. • Generar tablas y carga de datos en Hive. • Ejecutar queries y vistas. • Reconocer e implementar las principales operaciones básicas de optimización. • Implementar data pipelines en Hive.
5. Spark	Online: 1 Presencial: 2	• Conocer la arquitectura de Spark y sus principales APIs. • Entender el concepto de RDD. • Implementar ejemplos básicos. • Reconocer las diferencias entre el paradigma MapReduce vs Spark. • Reconocer e implementar las principales operaciones básicas de optimización. • Implementar ejemplos avanzados mediante PYSpark y casos de uso básicos de Machine Learning en Spark.

Módulo 6: Proyecto Data Science

Número de horas	• Sesión presencial: **18 horas** (6 sesiones de 3 horas cada una)
Descripción	El módulo de Proyecto está basado en aplicar los conocimientos y competencias adquiridas en los primeros cinco módulos en base a una problemática real.
Competencias	• Aplicar los conocimientos y resolver un problema real de la industria, ya sea mundo privado, gobierno o academia. • Implementar de forma eficaz el flujo de trabajo. • Presentar los resultados a una audiencia general, de manera eficaz y con óptimos resultados.

Alcance

UNIDAD	Nº SESIONES	ALCANCE
1. Diseño	Online: 0 Presencial: 2	• Identificar algún problema para solucionar, aplicando los conocimientos adquiridos en los primeros 3 módulos. • Diseñar e identificar los componentes necesarios para implementar el análisis (pregunta de investigación, diseño de hipótesis, elección de estrategia analítica).
2. Implementación	Online: 0 Presencial: 2	• Diseñar el flujo de trabajo en base a los componentes de los datos y la arquitectura computacional necesaria. • Diseñar e implementar el código necesario para un flujo de trabajo considerando preprocesamiento y manipulación de los datos, modelación y presentación de resultados.
3. Presentación	Online: 0 Presencial: 2	• Generar un reporte sobre los principales resultados de su trabajo, en base a la pregunta de investigación formulada. • Generar un pitch para presentar a audiencia general.



DESAFIO
LATAM

www.desafiolatam.com

inscripciones@desafiolatam.com

/DesafioLatam

/DesafioLatam