

## 1 Identificación Asignatura

<b>Nombre:</b>	Ingeniería de Datos		<b>Código:</b>	IN1032
<b>Carrera:</b>	Ingeniería Civil Industrial	<b>Unidad Académica:</b>	Ciencias Naturales y Tecnología	
<b>Ciclo Formativo:</b>		<b>Línea formativa:</b>		
<b>Semestre</b>	VII	<b>Tipo de actividad:</b>	Obligatoria	
<b>Nº SCT:</b>	6	<b>Horas Cronológicas Semanales</b>		
		<b>Presenciales:</b>	6	<b>Trabajo Autónomo:</b>
<b>Pre-requisitos</b>	IN1028 – Introducción a las TICA			

## 2 Propósito formativo

La asignatura de **Ingeniería de Datos** se diseña con el objetivo de dotar al estudiante de competencias críticas para identificar, gestionar y desarrollar soluciones innovadoras utilizando los vastos volúmenes de datos generados en diversos sectores industriales y sociales. En un entorno digital en constante evolución, es fundamental que los ingenieros no solo comprendan cómo almacenar y asegurar la calidad de los datos, sino también cómo extraer su valor intrínseco para impulsar decisiones estratégicas y operacionales.

El concepto de valorización de los datos, aunque no es nuevo, cobra especial relevancia en el marco de las cadenas de valor de datos. Estas cadenas ofrecen una visión integral del recorrido de los activos de datos, desde su adquisición hasta su explotación, facilitando la trazabilidad y maximización de su valor para las organizaciones. Los activos de datos, ya sean digitales, en papel o en otros formatos, constituyen recursos estratégicos cuya protección y gestión adecuada son prioritarias para las empresas modernas.

El curso se estructura en torno a las distintas fases del ciclo de vida de los datos, abarcando desde su generación y captura, hasta su administración, procesamiento y la aplicación de soluciones derivadas. Este enfoque modular asegura una comprensión profunda y práctica de la ingeniería de datos, permitiendo a los estudiantes enfrentarse a problemas de complejidad creciente. Iniciando con conceptos básicos de estadística y programación, los alumnos evolucionarán hacia la formulación de soluciones avanzadas, integrando tecnologías emergentes como la inteligencia artificial, aprendizaje automático, big data, computación en la nube y analítica de datos en tiempo real.

Esta asignatura es fundamental para la formación integral del estudiante, abordando el espectro completo del ciclo de vida de los datos a través de una pedagogía innovadora y multidisciplinaria. Desde las fases iniciales de recolección y tratamiento de datos hasta el almacenamiento y la programación avanzada, el curso equipa a los estudiantes con las habilidades esenciales para gestionar proyectos de ingeniería de datos con eficiencia y efectividad. Se enfatiza la importancia de comprender el contexto específico que motiva la recolección de datos, asegurando así que su posterior análisis y aprovechamiento generen un impacto tangible.

Además, el programa profundiza en la ciencia de datos, introduciendo a los estudiantes en los principios y técnicas del aprendizaje automático, tanto supervisado como no supervisado. Este enfoque práctico permite aplicar teorías y métodos en el análisis de conjuntos de datos reales a lo largo del curso, fomentando una comprensión profunda de cómo las tecnologías emergentes pueden ser utilizadas para extraer insights valiosos y resolver problemas complejos en diversas industrias. Integrando herramientas de vanguardia y metodologías actuales, como la inteligencia artificial, procesamiento de lenguaje natural y big data analytics, la asignatura prepara a los estudiantes no solo para enfrentar los desafíos del presente, sino también para innovar en el futuro de la ingeniería de datos.

Esta asignatura profundiza en los contenidos vistos previamente en Herramientas Informáticas, Programación I, Programación II, Introducción a las TICA, Probabilidades y Estadísticas Aplicadas, en cuanto aprovecha de manera general conceptos de desarrollo, gestión y puesta en marcha de soluciones que requieren de un enfoque multidisciplinar para la generación de valor a partir de los datos. Si bien no es una línea de profundización, permitirá a los estudiantes acercarse con mayor familiaridad y estar en la capacidad de asumir una postura técnico – científica de cara a proyectos relacionados con la generación, administración, análisis y propuesta de soluciones a partir de los datos.

### 3 Contribución al perfil de egreso

Esta asignatura contribuye a los siguientes desempeños o resultados de aprendizaje globales declarados en el Perfil de Egreso de la carrera:

- Demuestra dominio de las ciencias básicas, de energía y de las ciencias de la ingeniería y la capacidad de realizar análisis de tipo económico, políticos y sociales.
- Concibe soluciones a los problemas que surgen en las organizaciones.
- Diseña, selecciona y adapta desarrollos tecnológicos y científicos propios de la ingeniería industrial a los desafíos de las organizaciones.

### 4 Resultados de aprendizaje específicos (verbo + objeto + condición + finalidad)

- RA1. **Aplica** los principios de aprendizaje automático, incluyendo generalización, sobreajuste, y robustez, en la resolución de problemas de aprendizaje supervisado y no supervisado, con el fin de mejorar la precisión y eficacia de los modelos de datos.
- RA2. **Seleccionar** modelos de aprendizaje automático adecuados para situaciones específicas de negocio, considerando los supuestos subyacentes y limitaciones de cada modelo, para maximizar la relevancia y aplicabilidad de las soluciones de ingeniería de datos.
- RA3. **Analizar** necesidades empresariales mediante la formulación de preguntas críticas, para diseñar soluciones basadas en ingeniería de datos que aborden eficazmente estos requerimientos y contribuyan al éxito del modelo de negocio.
- RA4. **Distinguir** entre diferentes métodos y técnicas disponibles en repositorios y bibliotecas de código abierto para el modelado de problemas de negocio, con el propósito de seleccionar las herramientas más efectivas y eficientes para cada caso específico.
- RA5. **Implementar** fundamentos de aprendizaje supervisado y no supervisado para desarrollar y aplicar modelos en conjuntos de datos reales a lo largo del curso, con el objetivo de adquirir habilidades prácticas en el manejo, análisis, y interpretación de datos en contextos reales.

### 5 Unidades de Aprendizaje

#### 1. Introducción a Ingeniería de Datos

##### 1.1. Fundamentos de la Ingeniería de Datos

- Introducción al CRISP-DM
- Concepto de Data Frame
- Repositorios Git para gestión de proyectos de datos

##### 1.2. Manejo de Datos

- Datos abiertos y fuentes de datos (kaggle.com, pydataset)
- Lectura, limpieza y filtrado de datos
- Escritura y almacenamiento de datos

##### 1.3. Visualización de Datos

- Principios y fundamentos de visualización de datos
- Herramientas de visualización: matplotlib, seaborn
- Fundamentos a técnicas de visualización de datos

##### 1.4. Proyectos Prácticos y Herramientas

- Implementación de proyectos básicos
- Uso de Jupyter Notebooks para análisis de datos

#### 2. Conceptos Básicos de Estadística y Análisis Exploratorio

##### 2.1. Estadística Descriptiva

- Estadísticos básicos y gráficas
- Muestreo aleatorio y teorema central del límite
- 2.2. Inferencia Estadística**
  - Contrastes de hipótesis
  - Correlación y causalidad
- 2.3. Selección de Atributos**
  - Métodos y técnicas de selección de atributos
  - Importancia de variables en modelos predictivos
- 2.4. Proyecto Intermedio**
  - Aplicación de estadística en proyectos de datos
  - Avance de proyecto bajo metodología CRISP-DM

### 3. **Análisis de Regresión y Modelos Predictivos**

- 3.1. Regresión Lineal y Logística**
  - Regresión lineal simple y múltiple
  - Regresión logística para clasificación
- 3.2. Modelos Avanzados de Regresión**
  - Regresión polinomial
  - Árboles de regresión y Random Forest
- 3.3. Evaluación de Modelos**
  - Métricas de rendimiento
  - Validación cruzada
- 3.4. Proyectos de Regresión**
  - Desarrollo de modelos predictivos
  - Análisis de casos de estudio

### 4. **Clustering, Clasificación y Big Data**

- 4.1. Fundamentos de Clustering y Clasificación**
  - Concepto de distancia y métodos de enlace
  - K-medoides, clustering espectral, K-means
- 4.2. Modelos de Clasificación Avanzados**
  - K Nearest Neighbors (kNN)
  - Árboles de decisión y Random Forest
- 4.3. Introducción al Big Data**
  - Almacenamiento de grandes cantidades de datos
  - Ecosistema Hadoop: conceptos básicos y herramientas
- 4.4. Visualización Avanzada de Datos**
  - Análisis y visualización de datos con Tableau/Power BI
  - Conexión y manipulación de datos para visualización avanzada
  - Técnicas de visualización de datos geoespaciales, jerárquicos, temporales, textuales.

## 6 Recursos de Aprendizaje

### **Bibliografía:**

- B1. Downey, A. (2012). Think Python. " O'Reilly Media, Inc.". Versión en español gratuita y digital: <http://sourceforge.net/projects/httlcseifspa>
- B2. Han, J., Pei, J., & Kamber, M. (2011). **Data mining: concepts and techniques. Elsevier.**
- B3. Géron, A. (2019). **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.** O'Reilly Media.
- B4. McKinney, W. (2012). **Python for data analysis: Data wrangling with Pandas, NumPy, and IPython.** " O'Reilly Media, Inc."
- B5. Raschka, S., & Mirjalili, V. (2019). **Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2.** Packt Publishing Ltd.

### **Recursos materiales e infraestructura:**

- Acceso a Ucampus.
- Computadores debidamente equipados para utilizar lenguajes de alto nivel (por ej.: Python, Tableau, Power BI, otros).

## 7 Comportamiento y ética académica:

Se espera que, en el marco de sus actividades académicas y estudiantiles, los estudiantes demuestren un compromiso inquebrantable con los estándares de ética y honestidad académica que distinguen a nuestra comunidad universitaria. Este compromiso debe reflejarse en todas las acciones y trabajos realizados dentro del ámbito de la Universidad de Aysén, siguiendo rigurosamente las directrices establecidas en el Reglamento de Estudiantes, particularmente lo indicado en los artículos 23°, 24°, y 26°.

Además, es imperativo que todas las acciones se alineen con los reglamentos institucionales aplicables y las normativas internas vigentes, incluido el código de ética de la universidad. Estos documentos no solo rigen nuestros comportamientos cotidianos sino que también establecen las expectativas para nuestra convivencia y trabajo académico.

Cualquier infracción a estos principios de honestidad académica, detectada durante la realización, presentación, o entrega de cualquier actividad evaluativa del curso, acarreará consecuencias serias. Estas incluyen la suspensión inmediata de la actividad en cuestión y la asignación de la nota mínima posible (1.0) como reflejo de la gravedad del acto. Esta política se aplica de manera estricta y sin excepciones, reafirmando nuestro compromiso con una formación académica íntegra y responsable.

## Planificación del curso



### 8 Responsables

<b>Académico (s) Responsable (s) y equipo docente</b>	<b>Profesor:</b> Johnny Valencia C		
<b>Contacto</b>	<b>Correo:</b> <a href="mailto:johnny.valencia@uaysen.cl">johnny.valencia@uaysen.cl</a>		
<b>Año</b>	2024	<b>Periodo Académico</b>	Primer semestre
<b>Horario clases</b>	<b>Cátedra:</b> Lu 14.30 – 16.00 hrs*. Mi 14.30 – 16.00 hrs*.  <b>Sesión Prácticas:</b> Mi 16.15 – 17.45 hrs*.  *Presencial obligatoria	<b>Horario de atención estudiantes</b>	Contactar previamente al profesor mediante email.  Disponible Lu 16.00 – 18.00 Vi 14.30 – 16.00
<b>Sala / Campus</b>	Sala Virtual Ucampus – Google Sites		

### 9 Metodología de Trabajo:

En el curso se contemplan cuatro tipos de actividades docentes, las cuales se asocian a requerimientos de sala y al nivel de intervención del profesor/ayudante:

Actividad docente	Descripción	Intervención del profesor/ayudante	Requerimiento de sala
Exposición conceptual	El profesor introduce los fundamentos preliminares de programación y conceptos esenciales para el desarrollo de habilidades prácticas, con una participación activa en el aula tradicional o entornos virtuales.	Alta	Sala de clases UCampus Online UCampus Offline
Programación expositiva	Mediante la resolución de problemas específicos de ingeniería de datos, el profesor guía a los estudiantes a profundizar en el entendimiento de conceptos clave, empleando ejemplos prácticos en un entorno de sala de clases o plataformas online.	Alta	Sala de clases UCampus Online UCampus Offline
Programación tutorial	Esta modalidad incorpora pausas estratégicas en la exposición, permitiendo a los estudiantes completar tareas específicas bajo la supervisión del docente, favoreciendo el aprendizaje interactivo en laboratorios de computación o con uso de equipos personales	Media	Laboratorio de computación Computador persona
Actividad práctica / Programación autónoma	Fomenta la independencia de los estudiantes al enfrentar y resolver desafíos de ingeniería de datos, en grupos o individualmente, con mínima intervención docente, ideal para el trabajo en laboratorios de computación o mediante dispositivos personales.	Baja	Laboratorio de computación Computador persona

Durante el semestre, **se programarán evaluaciones menores** en cualquier momento, con el fin de revisar y reforzar los conocimientos adquiridos hasta la fecha. Estas evaluaciones podrán adoptar diversas formas:

- **Control (Quiz):** Evaluación, ya sea individual o en grupo, realizada en papel durante los primeros 45 minutos de la clase.
- **Trabajo en Clase:** Actividad evaluativa individual o grupal, llevada a cabo en computadora. Podrá ocupar los primeros 45 minutos o extenderse durante toda la sesión.

- **Entrega de Tareas:** Evaluación grupal realizada en computadora y fuera del horario de clase, promoviendo el trabajo colaborativo y la gestión del tiempo.

Adicionalmente, como componente de la calificación final, los estudiantes deberán presentar, en parejas, un proyecto que abarque el análisis, diseño e implementación de una solución tecnológica. Para esta tarea, se proporcionará un instructivo detallado que guiará a los estudiantes en su preparación y entrega, el desarrollo del proyecto se gestará durante todo el semestre, evidenciando una metodología clara de forma progresiva.

## 10 Evaluaciones:

Evaluación	Ponderaciones específicas	Ponderación nota presentación
Pruebas de cátedra	<ul style="list-style-type: none"> <li>• Prueba 1 (P1): 40%</li> <li>• Quiz 1 (Q1): 10%</li> <li>• Prueba 2 (P2): 40%</li> <li>• Quiz 2 (Q2): 10%</li> </ul>	60%
Evaluaciones menores	El promedio simple entre las notas consideradas corresponde al 100%.	40%

### Calificación final:

- Nota de presentación: 70%
- Examen Final: 30 % (Exposición de proyecto de aplicación)

### Requisito de Aprobación

- Asistencia: 65%
- Nota Final: 4,0
- La entrega de cualquiera de las evaluaciones indicadas en el programa de la asignatura por fuera del plazo definido serán evaluadas con nota mínima (1,0).
- Sólo para el caso de las evaluaciones presenciales, se permitirá comenzar con un atraso máximo de 15 minutos después de comenzada la misma. El tiempo de retraso no se podrá recuperar. Aquel/la estudiante que se presente a rendir la evaluación después de pasados los 15 minutos, se evaluará con nota mínima (1,0).
- En caso de inasistencia o no entrega en plazo de alguna evaluación, se podrán justificar mediante el mismo procedimiento antes descrito para la justificación de inasistencias actividades lectivas.
- Las clases teóricas tienen un 65% de asistencia mínima obligatoria, siguiendo el mínimo requerido por el Reglamento General de Estudios de Pregrado . La asistencia a las clases prácticas (laboratorios y terrenos) es del 100%. El no cumplimiento de estos porcentajes de asistencia será causal de reprobación de la asignatura.
- En casos debidamente justificados ante el Registro Académico, el/ la estudiante que no haya asistido a una salida a terreno o laboratorio tendrá derecho a rendir examen. Se considerarán debidamente justificadas las inasistencias ante el Registro Académico aquéllas que estén respaldadas con certificados médicos, laborales o algún documento validado por la Unidad de Acceso y Desarrollo Estudiantil. Las actividades de terrenos y laboratorios no podrán ser recuperadas.
- Se permitirá el ingreso posterior a la hora de inicio con un máximo de 15 minutos, siempre y cuando no sea una acción repetida por la/el estudiante (se aceptará máximo de 3 veces). Para las salidas a terreno, el tiempo de espera máximo será de 5 minutos. Para casos donde las actividades lectivas contemplen más de un módulo, el/la estudiante que no haya asistido al módulo anterior, podrá ingresar al comienzo del nuevo módulo.
- En caso de que ningún estudiante se presente a la actividad lectiva después de 15 minutos de comenzada, ésta se suspenderá. Los contenidos programados para dicha actividad se darán por dictados, será responsabilidad del estudiante ponerse al día con los contenidos de dicha clase. Los contenidos de dicha clase, y ejercicios, si así lo hubiera, serán enviados para ser realizados como trabajo autónomo.

## 11 Otros aspectos asociados al funcionamiento del curso:

- Toda la coordinación del curso (comunicaciones, actualización de notas, material, etc.) será realizada a través de UCampus.
- **Es deber del estudiante mantenerse informado de las noticias, avisos y material entregado por el profesor a través de estos medios, se sugiere instalar en su dispositivo móvil la aplicación de UCampus.**
- Recordar que los correos electrónicos serán respondidos en horario laboral (lunes a viernes de 9:00 a 18:00hs), no se responderán correos fuera de ese horario.
- Durante el desarrollo de las actividades lectivas, los teléfonos celulares deberán estar en silencio y guardados, a menos que el/la profesor/a específicamente requiera de estos equipos para la realización de su clase o durante algunos casos excepcionales conversados previamente con el/la docente a cargo.
- Las actividades lectivas se dictarán de forma presencial, salvo excepciones sujeto a contingencias presentes durante el transcurso de la asignatura.

## 12 Planificación de las actividades de enseñanza- aprendizaje y de evaluación

Semana	Resultado(s) de Aprendizaje (*)	Tema (Unidades de aprendizaje) y actividades	Recursos utilizados o lecturas (*)	Actividad(es)
1	RA1	<b>1.1. Fundamentos de la Ingeniería de Datos</b> <ul style="list-style-type: none"> <li>- Introducción al CRISP-DM</li> <li>- Concepto de Data Frame</li> <li>- Repositorios Git para gestión de proyectos de datos</li> </ul>	B1, B2	
2	RA1	<b>1.2. Manejo de Datos</b> <ul style="list-style-type: none"> <li>- Datos abiertos y fuentes de datos (kaggle.com, pydataset)</li> <li>- Lectura, limpieza y filtrado de datos</li> <li>- Escritura y almacenamiento de datos</li> </ul>	B1, B2	
3	RA1	<b>1.3. Visualización de Datos</b> <ul style="list-style-type: none"> <li>- Principios y fundamentos de visualización de datos</li> <li>- Herramientas de visualización: matplotlib, seaborn</li> <li>- Fundamentos a técnicas de visualización de datos</li> </ul>	B1, B2	
4	RA2	<b>1.4. Proyectos Prácticos y Herramientas</b> <ul style="list-style-type: none"> <li>- Implementación de proyectos básicos</li> <li>- Uso de Jupyter Notebooks para análisis de datos</li> </ul>	B1, B2	Prueba (P1) (3 de abril del 2024)
5	RA2, RA1	<b>2.1. Estadística Descriptiva</b> <ul style="list-style-type: none"> <li>- Estadísticos básicos y gráficas</li> </ul>	B1, B2	

		- Muestreo aleatorio y teorema central del límite		
6	RA4, RA2 RA1	<b>2.2. Inferencia Estadística</b> - Contrastes de hipótesis - Correlación y causalidad	B1, B2	
7	RA1, RA2	<b>2.3. Selección de Atributos</b> - Métodos y técnicas de selección de atributos - Importancia de variables en modelos predictivos	B1, B2	Quiz1 (Q1) (24 de abril del 2024)
8	RA4, RA2 RA1	<b>2.4. Proyecto Intermedio</b> - Aplicación de estadística en proyectos de datos - Avance de proyecto bajo metodología CRISP-DM	B1, B2	
9	RA4, RA2 RA1	<b>3.1. Regresión Lineal y Logística</b> - Regresión lineal simple y múltiple - Regresión logística para clasificación	B1, B2	
10	RA2, RA3, RA1	<b>3.2. Modelos Avanzados de Regresión</b> - Regresión polinomial - Árboles de regresión y Random Forest	B1, B2	
11	RA2, RA3, RA1	<b>3.3. Evaluación de Modelos</b> - Métricas de rendimiento - Validación cruzada	B1, B2	
12	RA3, RA1	<b>3.4. Proyectos de Regresión</b> - Desarrollo de modelos predictivos - Análisis de casos de estudio		Prueba P2(P2) (29 de mayo del 2024)
13	RA2, RA3, RA1	<b>4.1. Fundamentos de Clustering y Clasificación</b> - Concepto de distancia y métodos de enlace - K-medoides, clustering espectral, K-means	B1, B2	
14	RA2, RA3, RA1	<b>4.2. Modelos de Clasificación Avanzados</b> - K Nearest Neighbors (kNN) - Árboles de decisión y Random Forest	B1, B2	
15	RA2, RA3, RA1	<b>4.3. Introducción al Big Data</b> - Almacenamiento de grandes cantidades de datos - Ecosistema Hadoop: conceptos básicos y herramientas	B1, B2	Quiz 2 (Q2) (19 de junio del 2024)



16	RA2, RA3, RA1	<b>4.4. Visualización Avanzada de Datos</b> <ul style="list-style-type: none"> <li>- Análisis y visualización de datos con Tableau/Power BI</li> <li>- Conexión y manipulación de datos para visualización avanzada</li> <li>- Técnicas de visualización de datos geoespaciales, jerárquicos, temporales, textuales.</li> </ul>	B1, B2	
17		Periodo de exámenes		Exposición de un proyecto de aplicación (3 de Julio del 2024)
18		Periodo de exámenes		Exposición de un proyecto de aplicación

(\*) Para referencias de Resultados de Aprendizaje y Recursos, consultar el programa del curso.