INSTITUTO/S: Tecnología e Ingeniería

CARRERA/S: Tecnicatura Universitaria en Inteligencia Artificial

MATERIA: Fundamentos de Ciencias de Datos

NOMBRE DEL RESPONSABLE DE LA ASIGNATURA: Andrea Alejandra Rey

EQUIPO DOCENTE: Andrea Alejandra Rey

CUATRIMESTRE: 2do

AÑO: 2do

PROGRAMA N°: **(Aprob. por Cons. Directivo fecha XX)**

Instituto/s: Tecnología e Ingeniería
Carrera/s: Tecnicatura Universitaria en Inteligencia Artificial
Nombre de la materia: Fundamentos de Ciencia de Datos
Responsable de la asignatura y equipo docente: Andrea Alejandra Rey
Cuatrimestre y año: 2do cuat. del 2do año
Carga horaria semanal: 3 hs
Programa N°:
Código de la materia en SIU:

Fundamentos de Ciencias de Datos

1. Fundamentación

Debido a las grandes cantidades de datos disponibles en la actualidad, las empresas y las organizaciones se enfocan en la explotación de datos para obtener una ventaja competitiva, mejorar sus funciones y detectar problemas con el fin de solucionarlos. Antiguamente, se recurría a equipos de estadísticos/as, modeladores/as y analistas para explorar conjuntos de datos manualmente, pero el volumen y la variedad de datos han superado ampliamente la capacidad del análisis manual. Paralelamente, las computadoras se han vuelto mucho más poderosas, las redes se han vuelto omnipresentes, y se han desarrollado algoritmos que pueden conectar conjuntos de datos para un análisis más profundo. Es entonces cuando los principios de la Ciencia de Datos cobran una gran importancia. La materia pretende brindar ciertos conceptos básicos o principios fundamentales que subyacen a las técnicas para extraer conocimiento útil a partir de un conjunto de datos, y que se utilizan como base para muchos algoritmos de minería de datos, aprendizaje automático, aprendizaje profundo e inteligencia artificial.

2. Propósitos y/u objetivos

Propósitos

- Que los y las estudiantes incorporen a su práctica profesional el análisis exploratorio de datos.
- Extraer información a partir de un conjunto de datos: ¿son todos los datos válidos?, ¿cuáles son las mejores características que describen los datos?, ¿cuál es la manera más adecuada de visualizar el comportamiento de los datos?

Objetivos

Son objetivos de esta materia que los y las estudiantes:

- Reconozcan las habilidades requeridas para un/a científico/a de datos.
- Identifiquen los tipos de datos de interés.

- Adquieran herramientas para la manipulación de dataframes.
- Realicen un análisis exploratorio de los datos a partir de medidas estadísticas descriptivas clásicas.
- Presenten gráficamente la información obtenida a partir de los datos.
- Se inicien en modelos de regresión.
- Conozcan algunas técnicas para trabajar con datos faltantes.

3. Contenidos

3.1. Programa sintético

1. Ciencia de datos
2. R
3. Dataframes
4. Análisis exploratorio de datos usando medidas estadísticas descriptivas
5. Análisis exploratorio de datos mediante visualización gráfica
6. Regresión lineal simple
7. Imputación de datos

3.2. Unidades temáticas

1. Ciencia de datos
   Introducción. Tipos de datos. Big data. Proceso de datos. Características de un/a científico/a de datos. Aplicaciones de la Ciencia de Datos. Ciencia de datos e inteligencia artificial. Consideraciones éticas.

2. R
   Descarga e instalación del lenguaje R. Operaciones numéricas básicas. Asignaciones de variables. Operadores lógicos. Funciones. Arreglos. Operaciones escalares, vectoriales y matriciales. Lectura de archivos.

3. Dataframes
   Análisis univariado y multivariado. Manipulación de dataframes: combinación, selección y eliminación de datos. Limpieza de datos: definición y métodos.

4. Análisis exploratorio de datos usando medidas estadísticas descriptivas
   Medidas estadísticas: media, mediana, moda, rango, varianza, desvío estándar, cuartiles, percentiles. Valores atípicos. Frecuencias absoluta y relativa. Simetría y asimetría. Distribución Gaussiana. Covarianza y correlación.

5. Análisis exploratorio de datos mediante visualización gráfica
   Análisis univariado: Gráfico de barras. Gráfico circular o de torta. Gráfico de caja o boxplot. Histograma.

Análisis bivariado: Gráfico de barras agrupadas. Gráfico de barras segmentadas. Histogramas agrupados. Gráficos de densidad agrupados. Boxplots por grupos. Gráficos de violín. Gráficos de dispersión o scatterplots.
Series de tiempo: Gráfico de líneas. Tendencia. Estacionalidad. Ruido. Descomposición aditiva y multiplicativa.

6. Modelo de regresión lineal simple
   Modelo: definición. Recta de regresión lineal. Coeficiente de determinación. Predicciones.

7. Imputación de datos
   Análisis con datos completos. Análisis con los datos disponibles. Sustitución por la media. Imputación por vecinos más cercanos. Imputación por regresión.

3.3. Organización del contenido

El contenido se organiza a través de siete unidades, como se detalla a continuación.

La primera unidad tiene como objetivo introducir a los y las estudiantes en la Ciencia de Datos, presentando definiciones básicas, analizando los distintos tipos de datos y mostrando el esquema de proceso de datos. Del mismo modo, se describe el perfil de un/a científico/a de datos, se muestran campos de aplicación de la Ciencia de Datos y se la relaciona con la Inteligencia Artificial.

La segunda unidad pretende que los y las estudiantes se familiaricen con las sentencias básicas del lenguaje de programación R, ampliamente utilizado por profesionales del ámbito académico y por analistas de datos.

La tercera unidad tiene como propósito la incorporación de herramientas clásicas para la manipulación de estructuras de datos, conocidas como dataframes. Los temas desarrollados en esta unidad son centrales para el abordaje de los próximos contenidos.

Los próximas dos unidades engloban el análisis exploratorio de datos (EDA del inglés *Exploratory Data Analysis*), primer paso necesario en cualquier estudio basado en un conjunto de datos, con el fin de realizar una limpieza de datos, extraer características relevantes de los datos y elegir los modelos de ajuste o predicción, según el caso de interés. En particular, en la cuarta unidad se estudian medidas estadísticas descriptivas y en la quinta unidad se aplican técnicas gráficas, incluyendo un primer acercamiento al estudio de series de tiempo.

La sexta unidad tiene como objetivo introducir a los y las estudiantes a los modelos estadísticos a partir del algoritmo clásico de regresión linear simple, evaluando su rendimiento y aplicándolo para realizar predicciones.

Finalmente, la séptima unidad muestra algunas metodologías para trabajar con datos faltantes que van desde el descarte de registros hasta la imputación de datos empleando distintos métodos como la asignación de la media, por vecinos más cercanos y por regresión.

3.4. Bibliografía y recursos obligatorios

- Material teórico producido por la docente.
- Guías de ejercicios producidas por la docente.
- Trabajos prácticos producidos por la docente.

3.5. Bibliografía optativa:

- Bruce, P.; Bruce, A.; Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media.
- Cohen, Y.; Cohen, J. Y. (2008). *Statistics and data with R: An applied approach through examples*. John Wiley & Sons.
- Corea, F. (2018). *An introduction to data: Everything you need to know about AI, big data and data science*. Springer Cham.
- Fregly, C.; Barth, A. (2021). *Data science on AWS: Implementing end-to-end, continuous AI and machine learning pipelines*. O'Reilly Media, Inc.
- García, J.; Molina, J.; Berlanga, A.; Patricio, M.; Bustamante, A.; Padilla, W. (2018). *Ciencia de datos: Técnicas analíticas y aprendizaje estadístico*. Publicaciones Altaria, SL.
- Grus, J. (2019). *Data science from scratch: first principles with Python*. O'Reilly Media.
- Gutman, A. J.; Goldmeier, J. (2021). *Becoming a data head: How to think, speak, and understand data science, statistics, and machine learning*. John Wiley & Sons.
- Irizarry, R. A. (2019). *Introduction to data science: Data analysis and prediction algorithms with R*. CRC Press.
- Jägare, U. (2019). *Data science strategy for dummies*. John Wiley & Sons.

- Mertz, D. (2021). *Cleaning data for effective data science: Doing the other 80% of the work with Python, R, and command-line tools*. Packt Publishing Ltd.

- Nield, T. (2022). *Essential math for data science: Take control of your data with fundamental linear algebra, probability, and statistics*. O'Reilly.

- Park, A. (2021). *Data Science for Beginners: 4 Books in 1: Python programming, data analysis, machine learning. A complete overview to master the art of data science from scratch using Python for business*. Eureka Online Limited.

- Pierson, L. (2021). *Data science for dummies*. John Wiley & Sons.

- Provost, F.; Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

- Schutt, R.; O'Neil, C. (2014). *Doing data science: Straight talk from the frontline*. O'Reilly.

- Shah, C. (2020). *A hands-on introduction to data science*. Cambridge University Press.

- Varga, E. (2019). *Practical data science with Python 3: Synthesizing actionable insights from data*. Apress.

- Vaughan, D. (2020). *Analytical skills for AI and data science: Building skills for an AI-driven enterprise*. O'Reilly Media, Inc.

- Voulgaris, Z., & Bulut, Y. E. (2018). *AI for data science: Artificial intelligence frameworks and functionality for deep learning, optimization, and beyond*. Technics Publications, LLC.

4. Metodología de enseñanza

En las clases teóricas se presentan los métodos a utilizar, los fundamentos teóricos de los mismos, definiciones, algoritmos y ejemplos de aplicación. En las clases teóricas se promueve la participación de los alumnos mediante consultas o intervenciones vinculadas al contenido de la clase. El material de trabajo estará disponible en el campus virtual de la asignatura.

En las clases de consulta, los y las estudiantes asisten con dudas sobre los contenidos vistos en las clases teóricas o con dudas concretas acerca de la resolución da la guía de ejercicios o de los trabajos prácticos, tales como en la toma de una decisión, dificultades en la implementación (codificación en un lenguaje en particular), entre otras.

Plan de trabajo en el campus

El campus permite compartir el material teórico desarrollado en clase, los enunciados de las guías de ejercicios para realizar, ejemplos de código para que los y las estudiantes se familiaricen con la programación de los métodos, los enunciados de los trabajos prácticos y los archivos correspondientes a los conjuntos de datos necesarios para la resolución de las guías de ejercicios y de los trabajos prácticos.

Asimismo, se habilitarán foros de consultas sobre cada unidad temática propiciando un espacio común de debate e intercambio de ideas entre los y las estudiantes, el cual será monitoreado por la docente a cargo de la asignatura.

5. Actividades de investigación y extensión (si hubiera)

No aplica.

6. Evaluación y régimen de aprobación

6.1. Aprobación de la cursada

Para aprobar la cursada y obtener la condición de regularidad, el régimen académico establece que debe obtenerse una nota no inferior a cuatro (4) puntos. Todas las instancias evaluativas tienen una instancia de recuperación. Siempre que se realice una evaluación de carácter recuperatorio, la calificación obtenida en este examen reemplaza a la obtenida en el examen a recuperar, considerándose como definitiva a los efectos de la aprobación.

Por otra parte, el porcentaje de asistencia a las clases no puede ser inferior al 75%.

6.2. Aprobación de la materia

La materia puede aprobarse por promoción, evaluación integradora, examen final o libre.

Promoción directa: tal como lo establece el artículo 17° del Régimen Académico, para acceder a esta modalidad, el/la estudiante deberá aprobar la cursada de la materia con una nota no inferior a siete (7) puntos, no obteniendo en ninguna de las instancias de evaluación parcial menos de seis (6) puntos, sean evaluaciones parciales o recuperatorios. El promedio estricto resultante deberá ser una nota igual o superior a siete (7) sin mediar ningún redondeo.

Evaluación integradora: tal como lo establece el artículo 18° del Régimen Académico, podrán acceder a esta evaluación aquellos estudiantes que hayan aprobado la cursada con una nota de entre cuatro (4) y seis (6) puntos. La evaluación integradora tendrá lugar por única vez en el primer llamado a exámenes finales posterior al término de la cursada. Deberá tener lugar en el mismo día y horario de la cursada y será administrado, preferentemente, por el/la docente a cargo de la comisión. Se aprobará tal instancia con una nota igual o superior a cuatro

(4) puntos, significando la aprobación de la materia. La nota obtenida se promediará con la nota de la cursada.

Examen final: esta instancia está destinada a quienes opten por no rendir la evaluación integradora o hayan regularizado la materia en cuatrimestres anteriores. Se evalúa la totalidad de los contenidos del programa de la materia y se aprueba con una calificación igual o superior a cuatro (4) puntos. Esta nota no se promedia con la cursada.

6.3. Criterios de calificación

La calificación de cada evaluación se determinará en la escala 0 a 10, con la valoración indicada en la siguiente tabla.

Calificación	Valoración
0, 1, 2, 3	Insuficiente
4, 5	Regular
6, 7	Bueno
8, 9	Distinguido
10	Sobresaliente

8. Cronograma

Semana	Tema	Modalidad
1	Introducción a la Ciencia a datos	Presencial
2	Lenguaje R	Presencial
3	Dataframes	Presencial
4	Evaluación 1: Entrega del TP 1	Presencial
5	EDA: medidas estadísticas descriptivas	Presencial
6	EDA: covarianza y correlación	Presencial
7	EDA: visualización gráfica caso univariado	Presencial
8	EDA: visualización gráfica caso bivariado	Presencial
9	EDA: series de tiempo	Presencial
10	Clase de consultas para el TP 2	Presencial
11	Evaluación 2: Entrega del TP 2	Presencial
12	Regresión lineal simple	Presencial
13	Imputación de datos	Presencial
14	Clase de consultas para el TP 3	Presencial
15	Evaluación 3: Entrega del TP 3	Presencial
16	Clase de consultas para la instancia de recuperación	Presencial
17	Instancia de recuperación	Presencial