MSI603 – Proyecto Integrador: Ciencia de Datos

Mag. en Ingeniería Informática

22 de julio de 2022
Sesión 01

Identificación del Docente

Universidad Andrés Bello

IEEE Robotics & Automation Society | Chile Centro Section

IEEE Robotics & Automation Society
NEURO-ROBOTICS SYSTEMS
IEEE TECHNICAL COMMITTEE ON

ICDL-EPIROB
IEEE International Conference on Development and Learning and Epigenetic Robotics.

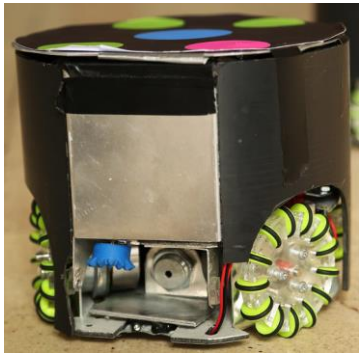
Director Ing. en Automatización y Robótica

Doctor en Ing. Informática

Magister en Ciencias de la Ing. Electrónica

Miguel Solis
miguel.solis@unab.cl

CRONE



Resultados de aprendizaje

- Identificar problemas, soluciones y sus implementaciones en las áreas de Ciencias de Datos e Ingeniería de Software, mediante la integración de diversas fuentes del ámbito académico y profesional.

- Diseñar soluciones computacionales innovadoras basadas en técnicas de Ciencias de Datos y/o Ingeniería de Software con impacto positivo en el proceso productivo de una organización pública o privada.

- Implementar soluciones computacionales basadas en técnicas de Ciencias de Datos y/o Ingeniería de Software acorde a un diseño debidamente validado.

- Comunicar efectivamente el diseño y la implementación de proyectos informáticos a la comunidad académica y profesional de manera escrita, verbal y visual.

Contenidos

Unidad I: Definición de proyecto

- Presentación de temas propuestos.
- Selección de temas propuestos.
- Definición de objetivos de temas propuestos.
- Comparación de soluciones existentes.

Unidad II: Análisis de datos

- Identificación y comprensión de conjuntos de datos.
- Análisis de datos.
- Identificación de outliers.
- Imputación de datos faltantes.
- Construcción de nuevas variables.

Contenidos

Unidad III: Implementación de modelos de ciencia de datos

- Particionamiento de los datos.
- Modelos de clasificación.
- Modelos de regresión.

Unidad IV: Validación y presentación de modelos de ciencia de datos

- Métricas de evaluación: MSE, MAE, RMSE, R2, Accuracy, Specificity, Sensitivity, Curva ROC.
- Técnicas de mejora de modelos. Manejo de sobreajuste.

Evaluación

- Informe de avance 1: 15%
- Informe de avance 2: 30%
- Informe de avance 3: 35%
- Exposición de proyecto: 20%

// Respecto a las presentaciones:

- Rúbrica disponible en Blackboard.
- Plantilla para informe en Blackboard.
- Proyecto grupal (3 integrantes).

Bibliografía

- Tavarez, A. (2021). Ciencia de Datos: Una Guía Práctica. Editorial Bienetre.

- Géron, A. (2020). Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow: Conceptos, herramientas y técnicas para construir sistemas inteligentes. Anaya Multimedia.

- Jones, H. (2019). Ciencia de los datos. Moliva AB.

Dataset 1: sueldos data scientists

work_year: año de pago
experience_level: experiencia durante ese año

- EN Entry-level / Junior
- MI Mid-level / Intermediate
- SE Senior-level / Expert
- EX Executive-level / Director

employment_type: tipo de empleo

- PT Part-time
- FT Full-time
- CT Contract
- FL Freelance

job_title: rol ejercido durante el año
salary: sueldo bruto anual
salary_currency: divisa en que se paga el sueldo
salaryinusd: sueldo en dólares
employee_residence: país de residencia durante ese año
remote_ratio: cantidad de trabajo remoto (0-50-100)
company_location: país del lugar de trabajo
company_size: cantidad promedio de trabajadores (S-M-L)

Dataset 2: supermercados

Store_ID: ID de la tienda en particular

Store_Area: área de la tienda en yardas al cuadrado

Items_Available: cantidad de distintos ítems disponibles en la tienda

DailyCustomerCount: número de clientes promedio

Store_Sales: ventas en dólares

Dataset 3: precios mundiales de petróleo

Country: país donde se analiza el precio

Daily Oil Consumption (Barrels): consumo diario en barriles

World Share: proporción del mundo

Yearly Gallons Per Capita: galones anuales per capita

Price per Gallon in USD: precio por galón en dólares

Price per Liter in USD: precio por litro en dólares

Price per Liter in PKR: precio por litro en rupias pakistaní

Dataset 4: precio del café

Date: fecha (por día)

Open: precio de apertura en ese día

High: precio máximo en ese día

Low: precio mínimo en ese día

Close: precio de cierre en ese día

Volume: volumen de kilos de café transados ese día

Currency: divisa del precio

Dataset 5: canales de YouTube

Rank: ranking del canal basado en cantidad de suscriptores

Youtuber: nombre oficial del canal

Subscribers: número de suscriptores que tiene el canal

Video views: número de vistas de todos los videos

Video count: número de videos que el canal ha subido

Category: género del contenido del canal

Started: año cuando el canal comenzó

Dataset 6: emisiones de CO2

Car: marca del automóvil

Model: modelo del automóvil

Volume: cantidad de automóviles de ese modelo y marca

Weight: peso promedio del automóvil de ese modelo y marca

CO2: nivel de co2 en partes por millón

Inscripción de dataset

Registrar en foro (hasta martes 26 de julio 23.59) :

- Nombre de integrantes
- Nombre del dataset escogido
- En caso de escoger dataset propio, incluir enlace de fuente