

PROGRAMA	Detección de anomalías o modelos de segmentación
ÁREA	Ingeniería
HORAS	12 horas, 6 por semana

Objetivo

El curso de detección de anomalías o modelos de segmentación ofrecido por el Programa de Educación Continua (PEC) de la Universidad Nacional de Colombia, desde el departamento de Ciencias económicas y el departamento de Ingeniería, tiene como objetivo proporcionar a los participantes metodologías propias de la estadística y el aprendizaje automático (Machine Learning) para la detección y el tratamiento de datos anómalos, aplicadas en el lenguaje python.

Contenido:

1. Introducción. Conceptos básicos.
2. Presentación de las técnicas tradicionales o más usuales.
 - a. Diagrama de caja
 - b. Análisis por Regresión: Análisis de residuales y Distancia de Cook: Leverage, Df_fit y residuales studentizados.
 - c. Random sample consensus
3. Presentación de técnicas del aprendizaje automático.
 - a. Vecino más cercano (Local Outlier Factor (LOF))
 - b. Árboles de decisión y ensambles (Isolation Forest)

Metodología

Si bien en las sesiones hay una parte destinada a la presentación de los temas por parte del docente, se tendrán talleres prácticos durante éstas con el objeto de que los estudiantes sean parte activa de su proceso de aprendizaje.

Evaluación

La sesión final desarrollará un estudio de caso en grupo en donde el participante en el curso deberá desarrollar las diversas técnicas aprendidas sobre un conjunto de datos.

Sesión 1.

Introducción al curso. Repaso de los paquetes numpy y pandas del lenguaje Python. Conceptos básicos sobre datos extremos y anómalos. Utilización de los diagramas de caja y los percentiles para detección de datos extremos de manera univariada.

Sesión 2.

Repaso de la regresión lineal y técnicas utilizadas para detección de datos anómalos en contextos multivariados. Demostración del Random sample consensus.

Sesión 3.

Introducción a las técnicas no supervisadas de aprendizaje automático para la detección de datos anómalos. Vecino más cercano y Local Outlier Factor (LOF), y Árboles de decisión y ensambles (Isolation Forest).

Sesión 4.

Sesión práctica. Análisis de caso.

Docentes:

José Fernando Zea Castro

Estadístico y Magíster en estadística de la Universidad Nacional de Colombia. Ha sido asesor del DNP, CEPAL, DANE y cuenta con más de 10 años de experiencia como consultor en diferentes entidades públicas y privadas. Con intereses en el diseño y desarrollo de encuestas, estimación de áreas pequeñas y Machine Learning.

Fernando López-Torrijos

Magister en Estadística Aplicada de la Universidad Santo Tomás de Bogotá, especialista en Sistemas de Información en la Organización de la Universidad de los Andes, especialista en Seguros y Seguridad Social de la Universidad de la Sabana. Ingeniero Industrial de la Universidad de los Andes. Ha trabajado en el sector público en la Administración Distrital de Bogotá, en el Ministerio de Educación y el Departamento Nacional de Planeación y en sector de las ONG en proyectos público-privados en el área de evaluación de proyectos. Fungió de actuario en Seguros de Vida Colpatria.