Minería de Datos con R

Sílabo

contacto

+51 (1) 282 9524
+51 9 9038 8434

info@perustat.com
http://www.perustat.com
fb://perustat

PeruStat
Analytics training, consulting & researching

fundamentación

Durante las últimas décadas se ha producido un desarrollo explosivo en las tecnologías de base de datos y la cantidad de datos que es recolectada. Esto ha creado una oportunidad sin precedentes para la Minería de Datos mediante el proceso de descubrimiento, ya sea supervisado o no, de información interesante y útil a partir de repositorios de datos disponibles.

La Minería de Datos está relacionada con el análisis, mayormente estadístico, de grandes conjuntos de datos con la finalidad de proporcionar ideas, patrones, modelos descriptivos y predictivos que permitan extraer y generar conocimiento para las organizaciones. Algunas de las tareas más comunes son la clasificación, agrupamiento, descubrimiento de reglas de asociación y patrones de respuestas, detección de anomalías, etc.

Dentro de los distintos programas estadísticos disponibles, el R proporciona una poderosa plataforma open source para la aplicación de la Minería de Datos, aunque por lo general el manejo de códigos y la programación es un reto para los analistas de datos que quieren utilizar esta herramienta. La libreria **Rattle** (R Analytical Tool To Learn Easily) adiciona una interfaz gráfica de usuario específicamente diseñada para facilitar la aplicación de la principales técnicas de Minería de Datos a los usuarios que no están acostumbrados al entorno de trabajo de este programa.

publico objetivo

Analistas e investigadores de mercado. Profesionales de marketing. Público en general que quiera adquirir conocimientos de Minería de Datos.

nivel

Introductorio

conocimientos previos

Conocimientos básicos en inferencia estadística y análisis de datos.

logros de aprendizaje

Al finalizar este curso, el participante conocerá los fundamentos de Minería de datos, aplicará las principales técnicas, así como analizará e interpretará los resultados obtenidos a través del programa estadístico R y la librería Rattle.

De manera específica el participante estará en capacidad de:

- Comprender el Proceso de Descubrimiento de Conocimiento en base de datos.
- Aplicar e interpretar adecuadamente las principales técnicas de Minería de datos.
- Comprender el desarrollo de los algoritmos de las principales técnicas de Minería de datos
- Usar el programa estadístico R, a través de las interfaz gráfica Rattle para el análisis e interpretación de las diferentes técnicas de Minería de Datos.

contenidos

UNIDAD 1: INTRODUCCIÓN

Sesión 1 **Conceptos** Básicos

- Breve historia de la Estadística y Minería de Datos. Conceptos básicos. Definición. Relación con otras disciplinas.
- Taxonomía de las técnicas de Minería de Datos: Tipos de modelos. Tipos de aprendizaje. Técnicas no supervisadas y supervisadas. Aplicaciones.
- Fases de la Minería de Datos. Descubrimiento de Conocimiento en Bases de datos (KDD).
- CRISP-DM: Estructura Básica. Fases.
- Herramientas de Minería de Datos. Instalación de R y de la librería Rattle.
- Primeros pasos con Rattle. Manejo de Datos.

UNIDAD 2: ANALÍTICA DESCRIPTIVA

Sesión 2 **Visualización** y Transformación de Datos

- Resumen de datos.
- Gráficas de distribuciones.
- Gráficas Interactivas.
- Transformación.
- Imputación.
- Reducción de la Dimensionalidad: Análisis de Componentes Principales.

Sesión 3 **Técnicas** de Segmentación

- Análisis de Conglomerados (Cluster): Definición. Requerimientos. Medición de la similaridad y distancias. Principales algoritmos.
- Conglomerados Jerárquicos
- Conglomerados no Jerárquicos: K-Medias. EWK (Entropy Weighted K-Means)

UNIDAD 3: ANALÍTICA PREDICTIVA

Sesión 4 **Regresión** Binaria

- Modelamiento predictivo: Conceptos básicos. Predicción numérica vs. clasificación. Precisión del modelo e interpretación. Balance entre la varianza y sesgo de un modelo predictivo.
- Modelos de Clasificación lineal y no lineal. Predicción y matrices de confusión.
- Clasificación binaria: Estimación del modelo de regresión logística binaria: Interpretación de los coeficientes. Validación del modelo.
- Modelo Probit.

Sesión 5 **Árboles** de Clasificación

- Árboles de Decisión: Representación. Partes de un Árbol de Decisión. Inducción y aprendizaje. Medidas de Selección de Atributos. Principales Algoritmos.
- Árboles de Clasificación y Regresión (CART). Construcción y poda del árbol.
- Árboles por inferencia condicional.

Sesión 6 **Evaluación** y Despliegue de un Modelo

- Evaluación: Matriz de Confusión. Curvas de Riesgo. Curvas ROC.
- Scoring.
- Predictive Model Markup Language (PMML): Exportación de modelos para su implementación.

metodología

La metodología del curso se basa en la aplicación de los conceptos teóricos en casos prácticos basados en datos reales. Cada sección del curso está motivada por un conjunto de datos en particular, de tal forma que el participante gane experiencia trabajando con una amplia variedad de fuentes de datos similares a los que usa en la realidad. Los contenidos del curso están estructurados en 6 sesiones con un total de 24 horas académicas.

evaluación y asistencia

Se otorgará un certificado a nombre de PeruStat Analytics S.A.C. que acredite la participación en el curso.

materiales

Material preparado por el equipo de capacitación con los contenidos del curso el cuál será entregado a los participantes en medios físicos (modalidad presencial) o digitales (modalidad virtual).

referencias

Clarke, B., Fokoue, E. y Zhang, H. (2009). Principles and Theory for Data Mining and Machine Learning. Springer Verlag.

Gareth, J., Witten, D., Hastie, T. y Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer Verlag.

Kuhn, M. y Johnson, K. (2013). Applied Predictive Modeling. Springer Verlag.

Larose, D.T. (2006). Data Mining Methods and Models. Wiley Interscience.

Ledolter, J. (2013). Data Mining and Business Analytics with R. John Wiley & Sons.

Nisbet, R., Elder IV, J. y Miner, G. (2009). Handbook of Statistical Analysis and Data Mining Applications. Academic Press.

Ohri, A. (2012). R for Business Analytics. Academic Press. Springer Verlag.

Putler, D. y Krider, R. E. (2012). Customer and Business Analytics: Applied Data Mining for Business Decision Making Using R. Chapman and Hall/CRC.

Williams, G. (2011). Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery. Springer Verlag.

Wu J. y Coggeshall, S. (2012). Foundations of Predictive Anaytics. Chapman and Hall/CRC.