PROMiDAT IBEROAMERICANO

Programa Iberoamericano de Formación en Minería de Datos

PROGRAMA

EXPERTO EN CIENCIA DE DATOS CON R

1. Introducción

Conviértase en un **Experto en Ciencia de Datos con R** y domine las herramientas que están dando forma al futuro de los negocios. Este programa está diseñado para capacitar a profesionales como usted en el análisis de datos, permitiéndole tomar decisiones informadas y transformar su organización. A través de este programa, aprenderá a interpretar, analizar y aplicar datos de manera eficiente, desbloqueando todo su potencial para generar impacto en su empresa.

El **Programa de Experto en Ciencia de Datos con R** se ha desarrollado como una formación personalizada y adaptable que permite a los estudiantes enfocarse en las áreas críticas que afectan directamente la toma de decisiones en su organización. Desde el análisis exploratorio hasta la predicción de inventarios y el modelado de series de tiempo, el enfoque práctico y basado en casos reales garantiza que lo aprendido pueda ser aplicado de inmediato en su entorno de trabajo.

10 años de experiencia enseñando 100% online respaldan este programa, el cual le permitirá compartir experiencias con profesionales de todo el mundo, y acceder a sesiones grabadas para mayor flexibilidad.

2. Perfil de Entrada de los Estudiantes

El Programa Experto en Ciencia de Datos con R está dirigido a personas de todas las áreas que estén interesadas en adquirir habilidades en el análisis de datos, sin requerir conocimientos previos en programación de computadoras.

Este programa es ideal para:

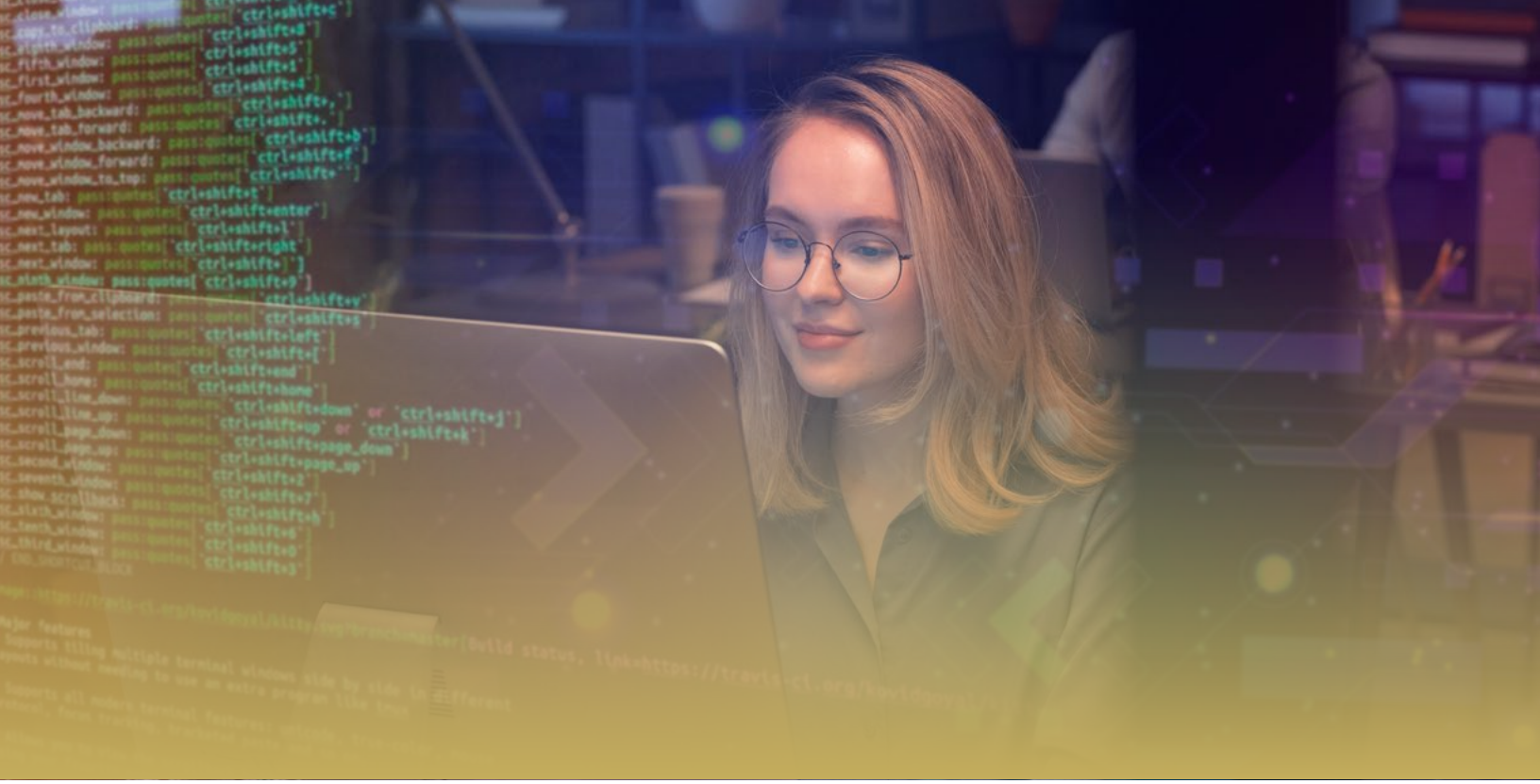
- **Analistas de riesgo y marketing:**
  que deseen usar datos para predecir tendencias y mejorar sus estrategias.

- **Profesionales de pronóstico y predicción:**
  que necesitan herramientas avanzadas para anticipar comportamientos y eventos.

- **Administradores de bases de datos:**
  que quieren optimizar la gestión de la información.

- **Estadísticos y economistas:**
  que buscan aplicar técnicas de minería de datos para generar insights clave en sus estudios.

- **Cualquier persona involucrada en la toma de decisiones basada en datos:**
  con interés en obtener habilidades que les permitan mejorar procesos, identificar oportunidades y generar valor en su empresa.



3. Objetivos Generales del Programa

El objetivo principal del Programa Experto en Ciencia de Datos con R es formar a profesionales capaces de:

- **Manipular y analizar datos:**
  Utilizando herramientas avanzadas de programación con R para aplicar técnicas de ciencia de datos en situaciones reales.

- **Generar modelos exploratorios, predictivos y de series de tiempo:**
  Que ayuden a prever comportamientos y eventos en distintos sectores empresariales.

- **Implementar soluciones prácticas:**
  Mediante proyectos aplicados a problemas reales, asegurando que los conocimientos adquiridos no solo sean teóricos, sino directamente utilizables en el día a día laboral.

- **Adaptarse a las necesidades empresariales:**
  Personalizando las técnicas y enfoques de análisis de datos para mejorar la toma de decisiones y optimizar los procesos en sus organizaciones.

4. Posibilidades de Empleo

Este programa abre las puertas a múltiples oportunidades laborales en el creciente campo de la ciencia de datos. Los egresados podrán desempeñarse en roles como:

- **Científico de Datos:**
  Responsable de diseñar, desarrollar y aplicar modelos predictivos y de análisis para resolver problemas complejos y generar valor en la empresa.

- **Analista de Datos:**
  Interpretando y analizando grandes volúmenes de datos para proporcionar insights útiles y guiar las decisiones estratégicas de las organizaciones.

- **Consultor en Ciencia de Datos:**
  Asesorando a empresas sobre cómo aprovechar los datos para mejorar sus procesos y aumentar la eficiencia organizacional.

- **Especialista en Business Intelligence (BI):**
  Desarrollando informes y reportes analíticos basados en la integración y análisis de datos para impulsar la mejora continua de las organizaciones.

5. Programa General

El Programa Experto en Ciencia de Datos con R tiene una duración total de 12 meses, con módulos que pueden ser tomados en cualquier orden y a través de nuestra plataforma online. Incluye clases en vivo por videoconferencia y acceso a un campus virtual 24/7, basado en Moodle, donde los estudiantes pueden interactuar con sus tutores y compañeros, además de acceder a grabaciones de las clases, materiales adicionales, software, y participar en debates a través de foros.

El programa está compuesto por los siguientes cursos:

1. **CD100:**
   Métodos Exploratorios en Ciencia de Datos
2. **CD103:**
   Métodos Predictivos en Ciencia de Datos
3. **CD106:**
   Métodos de Regresión en Ciencia de Datos
4. **CD200:**
   Métodos Avanzados en Ciencia de Datos.
5. **CD203:**
   Predicción de Inventarios y Series de Tiempo
6. **CD300:**
   Fundamentos de Programación en R
7. **CD303:**
   Programación Avanzada en R
8. **CD306:**
   Manipulación y Preparación de Datos
9. **CD309:**
   Visualización e Interpretación de Datos
10. **CD400:**
    Implementación de Proyectos en Ciencia de Datos
11. **CD900:**
    Proyecto Final de Graduación (2 meses)

6. Metodología

El "Programa Experto en Ciencia de Datos con R" tendrá una duración de 12 meses calendario, compartidos con estudiantes de todo el mundo y desarrollados 100% en la modalidad presencial-remoto.

Las clases se impartirán vía web en vivo desde nuestras oficinas, usando Zoom, mediante una vídeo conferencia, así el estudiante podrá seguir la lección desde cualquier lugar en donde se encuentre usando su computadora e incluso usando un Smartphone o una Tablet.
En estas vídeo conferencias participarán todos los estudiantes, quienes podrán hacer preguntas de forma oral o por medio del chat, en cualquier momento. Además, las videoconferencias quedarán grabadas y se subirán en el aula virtual, de modo que los estudiantes la puedan accesar tantas veces como lo necesiten, o en caso de que no pudieran asistir a la clase presencial-remoto.

12 Meses

Estudiantes de todo el mundo

Modalidad presencial-remoto

Nuestros cursos están basados en la teoría de la aplicación directa y práctica con casos reales de los conceptos aprendidos y tendrán una duración de 4 semanas.

Para esto se dispondrán de las siguientes herramientas:

Una vídeo conferencia semanal, la cual quedará grabada en Zoom, para que los alumnos la puedan acceder en cualquier momento, es decir, los estudiantes recibirán las clases en línea en tiempo real, pueden ver al tutor, seguir su presentación, hacerle preguntas y obtener respuestas en directo. Todas las asignaturas del programa cuentan con docencia en línea en tiempo real o en diferido, pueden ver las clases tantas veces como quiera según las necesidades del estudiante.

El alumno deberá realizar **un trabajo práctico semanal** que enviará a su tutor mediante el campus virtual, trabajos que serán evaluados por el tutor y remitidos al educando en un periodo corto de tiempo. A lo largo del curso contará con la ayuda del equipo de profesores que le orientará en su proceso de aprendizaje de manera continua. En resumen, nuestro estudiante nunca estará solo, siempre estará acompañado por el asistente, tutor y compañeros de curso.

El campus virtual, cuenta con herramientas de comunicación, tales como el correo electrónico, el foro de debate, la vídeo conferencia, entre otros que contribuirán a optimizar su proceso de enseñanza-aprendizaje.

El alumno tendrá acceso a un campus virtual en la plataforma Moodle donde podrá **descargar el material del curso:** Presentaciones, videoconferencia, tareas, software, material de apoyo entre otros. Todo lo necesario para llevar a cabo satisfactoriamente el curso en cuestión.

Se capacitará con **profesores con una amplia experiencia,** tanto docente como profesional.

7. Inversión

Horario de clases:
Martes 6:00 pm Costa Rica, la duración aproximada de la videoconferencia es de 2 horas.

Duración:
12 meses

Precio por curso:
$200 + 2% de IVA.

Precio Proyecto Final:
$300 +2% de IVA.

Inversión total:
$2300 + 2% de IVA

8. Información adicional:

Correo electrónico : info@promidat.com

Teléfono: +506 4030-1205

Web: www.promidat.com

WhatsApp: +506 8712-6978

Directo: +506 2271-0464

CD100
Métodos Exploratorios en Ciencia de Datos

Descripción

En este curso se presentarán los principales conceptos y métodos en Ciencia de Datos. El énfasis principal del curso será examinar dichos métodos desde un punto geométrico y de sus aplicaciones concretas. Se le dará especial importancia al uso de los conceptos de Ciencia de Datos en aplicaciones reales con bases de datos de gran tamaño, para esto se utilizarán los programas especializados en Ciencia de Datos como discoveR sobre la plataforma de software libre R y RStudio.

Objetivos

En este curso el estudiante será capaz de:

1. Entender la necesidad de la utilización de modelos, algoritmos, software especial para el descubrimiento de conocimiento en grandes volúmenes de datos.
2. Conocer la Metodología para el Desarrollo de Proyectos en Ciencia de Datos CRISP-DM.
3. Conocerá la metodología del ciclo de desarrollo usado para el descubrimiento del conocimiento en grandes bases datos (KDD – "Knowledge Discovery in Databases").
4. Entender las diferencias entre: estadística, análisis de datos, recuperación de la información, ML – "Machine Learning", Ciencia de datos y Ciencia de Datos.
5. Conocer los principales modelos, técnicas y algoritmos utilizados para descubrir el conocimiento en grandes volúmenes de datos.
6. Utilizar el discoveR sobre la plataforma R para analizar ejemplos con datos reales.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.
3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Ciencia de Datos que involucren segmetación de carteras de clientes.

Contenido

1. **Conceptos de la Ciencia de Datos.**
   - **a.** Definiciones básicas en Ciencia de Datos.
   - **b.** Instalación de la Plataforma R y RStudio.
   - **c.** Instalación del paquete discoveR.
2. **Análisis Exploratorio de Datos.**
   - **a.** Tipos de variables.
   - **b.** Estadísticas básicas y matriz de correlaciones.
   - **c.** Tablas de datos y datos atípicos.
   - **d.** Aplicaciones en casos reales con el paquete discoveR sobre la plataforma R.
3. **Métodos de Reducción de la Dimensión.**
   - **a.** Análisis en Componentes Principales – ACP (PCA, Karhunen-Loeve o K-L Method).
   - **b.** Plano principal.
   - **c.** Círculo de correlaciones.
   - **d.** Dualidad y sobre-posición de gráficos.
   - **e.** Análisis Factorial de Correspondencias Múltiples.
   - **f.** Aplicaciones en casos reales con el paquete discoveR

4. **Clustering Jerárquico Aglomerativo.**
   - **a.** ¿Qué es "cluster analysis"?
   - **b.** Clustering Jerárquica Aglomerativa.
   - **c.** Distancias y matrices de distancias.
   - **d.** Agregaciones.
   - **e.** Jerarquías binarias.
   - **f.** Jerarquías Binarias sobre las Componentes Principales.
   - **g.** Aplicaciones en casos reales con discoveR.
5. **Método de k-medias (k-means).**
   - **a.** Inercia total, inercia inter-clases e inercia intra-clases.
   - **b.** Teorema de Fisher.
   - **c.** Problema combinatorio.
   - **d.** Método de Forgy.
   - **e.** Método de las nubes dinámicas.
   - **f.** Aplicaciones en casos reales con R y el paquete discoveR.

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] Analyses factorielles simples - Xavier Bry - Librairie Eyrolles. url: https://www.eyrolles.com/Sciences/Livre/analyses-factorielles-simples-9782717828597/ (visited on10/17/2022).

[2] W. John Braun and Duncan J. Murdoch. A First Course in Statistical Programming with R. Google-Books-ID: NzorEAAAQBAJ. Cambridge University Press, May 20, 2021.281 pp. isbn: 978-1-108-99514-6.

[3] Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 3rd Edition | Wiley. Wiley.com. url: https://www.wiley.com/en-us/Data+Mining+Tec hniques%3A+For+Marketing%2C+Sales%2C+and+Customer+Relationship+Management%2C+3rd+Edition-p-9781118087459 (visited on 10/17/2022).

[4] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY: Springer, 2001. isbn: 978-1-4899-0519-2 978-0-387-21606-5. doi: 10.1007/978-0-387-21606-5. url: http://link.springer.com/10.1007/978-0-387-21606-5 (visited on 10/17/2022).

[5] Michel Jambu. "Introduction au data mining". In: (1999).

[6] Boris Mirkin and Boris Mirkin. Clustering for Data Mining: A Data Recovery Approach. New York: Chapman and Hall/CRC, Apr. 28, 2005. 296 pp. isbn: 978-0-429-13754-9.doi: 10.1201/9781420034912.

[7] R: The R Project for Statistical Computing. url: https://www.r-project.org/.

[8] O Rodríguez, PJF Groenen S Winsberg, and E Diday. "I-Scal: Symbolic Multidimensional Scaling of Interval Dissimilarities". In: COMPUTATIONAL STATISTICS & DATA ANALYSIS the Official Journal of the International Association for Statistical Computing, London (2006).

CD103
Métodos Predictivos en Ciencia de Datos

Descripción

En este curso se presentarán los principales métodos en Ciencia de Datos, especialmente enfocados en métodos predictivos, conocidos también como métodos de aprendizaje supervisado. El énfasis principal del curso será examinar dichos métodos desde un punto de vista algorítmico y de sus aplicaciones en casos reales. Se le dará especial importancia al uso de los conceptos de Ciencia de Datos en aplicaciones reales con bases de datos de gran tamaño, para esto se utilizarán los programas especializados en Ciencia de Datos, como son la plataforma de desarrollo R y el paquete predictoR.

Objetivos

En este curso el estudiante será capaz de:

1. Comprender la diferencia entre modelos de aprendizaje supervisado (minería predictiva) y modelos de aprendizaje no supervisado (minería descriptiva).
2. Comprender la diferencia entre bases de datos de aprendizaje y bases de datos de "testing".
3. Comprender la necesidad de la utilización de modelos, algoritmos, software para predecir el comportamiento futuro.
4. Conocer los principales modelos predictivos, técnicas y algoritmos utilizados para predecir conductas a partir de grandes volúmenes de datos históricos.
5. Utilizar la plataforma R y el paquete predictoR para analizar y desarrollar ejemplos con datos reales.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.
3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Ciencia de Datos que involucren predicción utilizando modelos predictivos.

Contenido

1. **Conceptos de la Analítica Predictiva.**
   - **a.** Conceptos y diferencias entre aprendizaje supervisado y aprendizaje no supervisado.
   - **b.** Diseño de bases de datos de aprendizaje.
   - **c.** Diseño de bases de datos de testing.
   - **d.** Variables cuantitativas y variables cualitativas.
   - **e.** ¿Cómo evaluar la calidad de un modelo predictivos?
   - **f.** Cálculo de la Matriz de confusión e índices de calidad.
   - **g.** Curvas ROC.
   - **h.** Aplicación con datos reales con predictoR.Análisis Exploratorio de Datos.
2. **Método de los K vecinos más cercanos.**
   - **a.** Estructura General del método.
   - **b.** El mejor valor de K.
   - **c.** Algoritmo de Aprendizaje.
   - **d.** Aplicación con datos reales con predictoR.

3. **Máquinas Vectoriales de Soporte.**
   - **a.** Hiperplano de separación de las clases.
   - **b.** Vectores de soporte.
   - **c.** Función discriminante lineal.
   - **d.** ¿Cómo resolver un Problema Optimización?
   - **e.** MVS no linealmente separables.
   - **f.** Núcleos en Máquinas Vectoriales de Soporte.
   - **g.** Aplicación con datos reales con predictoR.
4. **Árboles de Decisión (Método CART).**
   - **a.** Algoritmos ID3, C4.5, C5.0 y CART.
   - **b.** Árboles de auto-regresión.
   - **c.** Aplicación con datos reales con predictoR.
5. **Métodos de consenso y de Potenciación.**
   - **a.** Métodos de Consenso (Bagging).
   - **b.** Bosques Aleatorios (Random forests).
   - **c.** Métodos de impulso (Boosting).
   - **d.** Métodos de Potenciación (ADA Boosting).
   - **e.** Aplicación con datos reales con predictoR.

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] Esteban Alfaro, Matias Gamez, and Noelia García. "adabag: An R Package for Classification with Boosting and Bagging". In: Journal of Statistical Software 54 (Sept. 3, 2013), pp. 1–35. issn: 1548-7660. doi: 10.18637/jss.v054.i02. url: https://doi.org/10.18637/jss.v054.i02 (visited on 10/19/2022).

[2] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: TEST 25.2 (June 1, 2016), pp. 197–227. issn: 1863-8260. doi: 10.1007/s11749-016-0481-7. url: https://doi.org/10.1007/s11749-016-0481-7 (visited on 10/19/2022).

[3] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY: Springer, 2001. isbn: 978-1-4899-0519-2 978-0-387-21606-5. doi: 10.1007/978-0-387-21606-5. url: http://link.springer.com/10.1007/978-0-387-21606-5 (visited on 10/19/2022).

[4] Alexandros Karatzoglou, David Meyer, and Kurt Hornik. "Support Vector Machines in R". In: Journal of Statistical Software 15 (Apr. 6, 2006), pp. 1–28. issn: 1548-7660. doi:10.18637/jss.v015.i09. url: https://doi.org/10.18637/jss.v015.i09.

[5] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: 2 (2002), p. 5.

[6] Wei-Yin Loh. "Classification and regression trees". In: WIREs Data Mining and Knowledge Discovery 1.1 (2011). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.8,pp. 14–23. issn: 1942-4795. doi: 10.1002/widm.8. url: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.8.

[7] R: The R Project for Statistical Computing. url: https://www.r-project.org/.

CD106
Métodos de Regresión en Ciencia de Datos

Descripción

En este curso se presentarán los principales métodos en Ciencia de Datos, especialmente enfocados en métodos de Regresión. El énfasis principal del curso será examinar dichos métodos desde un punto de vista algorítmico y de sus aplicaciones en casos reales. Se le dará especial importancia al uso de los conceptos de Ciencia de Datos en aplicaciones reales con bases de datos de gran tamaño, para esto se utilizarán los programas especializados en Ciencia de Datos, como son la plataforma de desarrollo R y el paquete regressoR.

Objetivos

En este curso el estudiante será capaz de:

1. Comprender la diferencia entre modelos de clasificación y modelos de regresión.
2. Comprender la diferencia y necesidad de usar bases de datos de aprendizaje y bases de datos de testing.
3. Comprender la necesidad de la utilización de modelos, algoritmos, software para predecir el comportamiento futuro.
4. Conocer los principales modelos de regresión, técnicas y algoritmos utilizados para predecir conductas a partir de grandes volúmenes de datos históricos.
5. 5. Utilizar la plataforma R y el paquete regressoR para analizar y desarrollar ejemplos con datos reales.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.
3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Ciencia de Datos que involucren la predicción de una variable numérica utilizando modelos de Regresión.

Contenido

1. **Conceptos Básicos de Regresión.**
   - **a.** Conceptos y diferencias entre clasificación y regresión
   - **b.** Diseño de bases de datos de aprendizaje.
   - **c.** Diseño de bases de datos de testing.
   - **d.** Manipulación de variables cuantitativas y variables cualitativas.
   - **e.** ¿Cómo evaluar la calidad de un modelo un modelo de regresión?
   - **f.** Índices de error. Error cuadrático medio, error absoluto medio, correlación y otros.
2. **Regresión Lineal Clásica.**
   - **a.** Regresión Lineal simple y múltiple.
   - **b.** Interpretación de coeficientes.
   - **c.** Aplicaciones en casos reales con el paquete regressoR.
3. **Método de Regresión Lineal Penalizada.**
   - **a.** Regresión LASSO.
   - **b.** Regresión RIDGE.
   - **c.** Regresión combinando métodos de reducción de la dimensión ACP y MCP.
   - **d.** Aplicaciones en casos reales con el paquete regressoR.

4. **Árboles de Decisión para Regresión – Método CART.**
   - **a.** Árboles de Regresión.
   - **b.** Bosques de Regresión.
   - **c.** Boosting para Regresión.
   - **d.** Aplicaciones en casos reales con el paquete regressoR.
5. **Máquinas de Soporte Vectorial para Regresión.**
   - **a.** Máquinas de Soporte Vectorial en problemas regresión.
   - **b.** Truco del Kernel.
   - **c.** Aplicaciones en casos reales con el paquete regressoR.
6. **Deep Learning en Problemas de Regresión.**
   - **a.** Redes Neuronales y su estructura. Neuronas, capas, funciones de activación.
   - **b.** Redes Neuronales para problemas predictivos.
   - **c.** Redes Neuronales y Deep Learning para problemas de regresión.
   - **d.** Aplicaciones en casos reales con el paquete regressoR.

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] Gérard Biau and Erwan Scornet. "A random forest guided tour". In: TEST 25.2 (June 1, 2016), pp. 197–227. issn: 1863-8260. doi: 10.1007/s11749-016-0481-7. url: https://doi.org/10.1007/s11749-016-0481-7 (visited on 10/19/2022).

[2] Brad Boehmke and Brandon M. Greenwell. Hands-On Machine Learning with R. New York: Chapman and Hall/CRC, Nov. 14, 2019. 488 pp. isbn: 978-0-367-81637-7. doi:10.1201/9780367816377.

[3] John Fox and Sanford Weisberg. An R Companion to Applied Regression. Google-Books-ID: uPNrDwAAQBAJ. SAGE Publications, Sept. 27, 2018. 608 pp. isbn: 978-1-5443-3648-0.

[4] Trevor Hastie, Jerome Friedman, and Robert Tibshirani. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY: Springer, 2001. isbn: 978-1-4899-0519-2 978-0-387-21606-5. doi: 10.1007/978-0-387-21606-5. url: http://link.springer.com/10.1007/978-0-387-21606-5 (visited on 10/19/2022).

[5] Alexandros Karatzoglou, David Meyer, and Kurt Hornik. "Support Vector Machines inR". In: Journal of Statistical Software 15 (Apr. 6, 2006), pp. 1–28. issn: 1548-7660. doi:10.18637/jss.v015.i09. url: https://doi.org/10.18637/jss.v015.i09.

[6] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In:2 (2002), p. 5.

[7] Wei-Yin Loh. "Classification and regression trees". In: WIREs Data Mining and Knowledge Discovery 1.1 (2011). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.8,pp. 14–23. issn: 1942-4795. doi: 10.1002/widm.8. url: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.8.

[8] R: The R Project for Statistical Computing. url: https://www.r-project.org/.

[9] Simon Sheather. A Modern Approach to Regression with R. Google-Books-ID: IZ6Ra0efSdsC. Springer Science & Business Media, Feb. 27, 2009. 398 pp. isbn: 978-0-387-09608-7.

CD200
Métodos Avanzados en Ciencia de Datos

Descripción

En este curso se estudiarán en detalle las técnicas de Validación Cruzada (Cross-Validation) y Remuestreo (boostrapping) con el objetivo de calibrar y seleccionar el mejor método de Ciencia de Datos para un problema y juego de datos dado. Se estudiará como programar y automatizar la Validación Cruzada (Cross-Validation) y Remuestreo (boostrapping) en el lenguaje R. Los ejemplos de este curso estarán motivados por problemas reales en el campo. Por lo tanto, los estudiantes adquieren conocimientos de muchas herramientas diferentes que pueden combinarse para resolver problemas reales.

Objetivos

En este curso el estudiante será capaz de:

1. Usar el paquete predictoR para Validación Cruzada (Cross-Validation) y Remuestreo (bootsrapping).
2. Aplicar adecuadamente una Validación Cruzada y un Remuestreo.
3. Calibrar adecuadamente tanto métodos exploratorios como métodos predictivos en Ciencia de Datos.
4. Seleccionar el mejor modelo predictivo dado un conjunto de datos.
5. Instalar y utilizar paquetes avanzados en R.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.

3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Ciencia de Datos que involucren calibración y selección de métodos exploratorios y predictivos en código R.

Contenido

1. **Métodos Bayesianos**
   - **a.** Método de Bayes.
   - **b.** Análisis Discriminante lineal y cuadrático.
2. **Probabilidad de Corte y Curva ROC.**
   - **a.** Probabilidad de corte.
   - **b**. Curva ROC.
3. **Validación Cruzada (cross-validation).**
   - **a.** Enfoque: "tabla de aprendizaje y tabla de testing" (the validation test approach).
   - **b.** Validación cruzada dejando uno fuera (Leave-one-out cross-validation - LOOCV).
   - **c.** Validación cruzada usando K grupos (K-fold cross-validation).
4. **Calibración y Selección de Métodos.**
   - **a.** Calibración de Métodos Exploratorios (descriptivos).
   - **b.** Calibración de Métodos Predictivos.
   - **c.** Seleccionando el mejor método predictivo.

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] Hervé Abdi and Dominique Valentin. "Multiple Correspondence Analysis". In: Multiple Correspondence Analysis (), p. 13.

[2] Michael Greenacre and Jorg Blasius, eds. Multiple Correspondence Analysis and Related Methods. New York: Chapman and Hall/CRC, June 22, 2006. 608 pp. isbn: 978-0-429-14196-6. doi: 10.1201/9781420011319.

[3] H. Teil. "Correspondence factor analysis: An outline of its method". In: Journal of the International Association for Mathematical Geology 7.1 (Feb. 1, 1975), pp. 3–12. issn:1573-8868. doi: 10.1007/BF02080630. url: https://doi.org/10.1007/BF02080630 (visited on 10/19/2022).

[4] The 'K' in K-fold Cross Validation. url: https://arpi.unipi.it/handle/11568/962587 (visited on 10/19/2022).

[5] R. Vidal, Yi Ma, and S. Sastry. "Generalized principal component analysis (GPCA)". In: IEEE Transactions on Pattern Analysis and Machine Intelligence 27.12 (Dec. 2005). Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1945–1959. issn: 1939-3539. doi: 10.1109/TPAMI.2005.244.

[6] Tzu-Tsung Wong. "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation". In: Pattern Recognition 48.9 (Sept. 1, 2015), pp. 2839–2846. issn: 0031-3203. doi: 10.1016/j.patcog.2015.03.009. url: https://www.sciencedirect.com/science/article/pii/S0031320315000989 (visited on 10/19/2022).

CD203
Predicción de Inventarios Series de Tiempo

Descripción

En este curso se presentarán los principales conceptos, algoritmos y métodos en Series de Tiempo. El énfasis principal del curso será examinar dichos métodos desde un punto geométrico y de sus aplicaciones concretas. Se le dará especial importancia al uso de los conceptos en aplicaciones reales con bases de datos de gran tamaño, para esto se utilizarán algunos paquetes especializados sobre la plataforma de software libre R.

Objetivos

En este curso el estudiante será capaz de:

1. Utilizar el paquete forecasteR en R para trabajar con Series de Tiempo.
2. Aprender procesos y técnicas para preparar y visualizar Series de Tiempo en forecasteR.
3. Entender la necesidad de la utilización de modelos, algoritmos, software especial para la predicción en Series de Tiempo.
4. Aprender las técnicas estadísticas aplicables a datos tipo series de tiempo para preparar pronósticos.
5. Aprender los métodos basados en regresión, suavizado exponencial, Método de Holt Winters, métodos ARIMA y Redes Neuronales.
6. Utilizar forecasteR para analizar ejemplos con datos reales de Series de Tiempo.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.

2. Trabajos prácticos semanales.
3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Minería de Datos que involucren manipulación, corrección, calibración y predicción de modelos en Series de Tiempo.

Contenido

1. **Análisis de Series de Tiempo**
   - **a.** ¿Qué es una serie de tiempo?
   - **b.** Instalación del paquete forecasteR.
   - **c.** Preparación de los archivos para que sean leídos por forecasteR.
   - **d.** Corrección de fechas.
   - **e.** Estacionalidad.
   - **f.** Visualización de una serie de tiempo.
   - **g.** Cálculo de la normalidad de las series de tiempo.
   - **h.** Series de Fourier y Regresión.
   - **i.** Filtrado Lineal Descomposición de las series de tiempo: tendencia, ciclos y estacionalidades.
   - **j.** Periodograma.
   - **k.** Pronóstico y medición de errores.
2. **Ajuste exponencial**
   - **a.** Suavizado Exponencial.
   - **b.** Modelado de Holt Winters.
   - **c.** Calibración del modelo de Holt Winters.
   - **d.** Predicción mediante los modelos de Holt Winters.
3. **Métodos ARIMA**
   - **a.** Análisis de las autocorrelaciones y autocorrelaciones parciales.
   - **b.** Análsis del periodograma para los modelos ARIMA.

c. Estimación de parámetros en los modelos ARIMA.

d. Modelado de ARIMA.

e. Predicción mediante los modelos ARIMA.

4. **Métodos Basados en Redes Neuronales y Deep Learning.**

a. Redes Neuronales Recurrentes (RNN).

b. Redes Neuronales de memoria a corto y largo plazo (LSTM).

c. Modelado de métodos basados en Redes Neuronales.

d. Predicción mediante los modelos basados en Redes Neuronales.

e. Uso de reglas.

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] Peter J. Brockwell and Richard A. Davis. Time Series: Theory and Methods. Springer Series in Statistics. New York, NY: Springer, 1991. isbn: 978-1-4419-0319-8 978-1-4419-0320-4. doi: 10.1007/978-1-4419-0320-4. url: http://link.springer.com/10.1007/978-1-4419-0320-4 (visited on 10/05/2022).

[2] John Chambers. Software for Data Analysis. Statistics and Computing. New York, NY: Springer, 2008. isbn: 978-0-387-75935-7 978-0-387-75936-4. doi: 10.1007/978-0-387-7593 6-4. url: http://link.springer.com/10.1007/978-0-387-75936-4 (visited on 10/05/2022).

[3] Avril Coghlan. "A Little Book of R For Time Series". en. In: (), p. 81.

[4] John Cristian Borges Gamboa. Deep Learning for Time-Series Analysis. arXiv:1701.01887 [cs]. Jan. 2017. doi: 10.48550/arXiv.1701.01887. url: http://arxiv.org/abs/1701.01887 (visited on 10/05/2022).

[5] Rami Krispin. Hands-On Time Series Analysis with R: Perform time series analysis and forecasting using R. en. Google-Books-ID: tTmbDwAAQBAJ. Packt Publishing Ltd, May 2019. isbn: 978-1-78862-404-6.

[6] Andrew V. Metcalfe and Paul S.P. Cowpertwait. Introductory Time Series with R. en. New York, NY: Springer New York, 2009. isbn: 978-0-387-88697-8 978-0-387-88698-5. doi: 10.1007/978-0-387-88698-5. url: http://link.springer.com/10.1007/978-0-387-88698-5 (visited on 10/05/2022).

[7] Wayne A. Woodward, Henry L. Gray, and Alan C. Elliott. Applied Time Series Analysis with R. 2nd ed. Boca Raton: CRC Press, Jan. 2017. isbn: 978-1-315-16114-3. doi: 10.1201/9781315161143.4

CD300
Fundamentos de Programación en R

Descripción

En este curso se presentarán desde cero los principales conceptos de programación y su aplicación en lenguaje R, enfocado al ámbito de la ciencia de datos.

Objetivos

En este curso el estudiante será capaz de:

1. Crear scripts de R.
2. Comprender el proceso de importación de datos en formato csv y xlsx.
3. Utilizar funciones estadísticas de R.
4. Crear soluciones a problemas personalizados creando sus propias funciones.
5. Utilizar Rstudio como herramienta para el desarrollo de análisis de datos.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.
3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar scripts en lenguaje R que permiten un análisis de datos básico.

Contenido

1. **Instalación y preparación del entorno de trabajo.**
   - **a.** Instalar RStudio.
   - **b.** Instalar R.
   - **c.** Instalaciones complementarias para Windows.
   - **d.** Instalaciones complementarias para MacOS.
2. **Introducción a los diagramas de flujo estructurados**
   - **a.** Esquema general de especificación.
   - **b.** Modulación y bloque fundamental.
   - **c.** Seguimiento de un diagrama de flujo.
   - **d.** Estructuras fundamentales en un diagrama de flujo.
   - **e.** El bloque de dos opciones (sí).
     - ii. El bloque de opciones múltiples (case).
     - iii. Condiciones compuestas.
     - iv. El bloque de entrada condicionada (mientras).
     - v. El bloque de entrada asegurada (hasta que).
3. **Tipos de datos.**
   - **a.** Vectores Atómicos.
   - **b.** Matrices.
   - **c.** Listas.
   - **d.** Tablas de datos.
4. **Lectura de datos.**
   - **a.** Lectura de datos csv y xlsx.
   - **b.** Tipos de datos y validación de archivos.
   - **c.** Modificar tipos de datos de un archivo.
   - **d.** Predicción mediante los modelos de Holt Winters.

5. **Estructuras de control.**
   - **a.** IF - ELSE.
   - **b.** Ciclos FOR.
   - **c.** Ciclos While.
6. **Funciones.**
   - **a.** Estructura de una función.
   - **b.** Buenas prácticas para el desarrollo de funciones.
   - **c.** Manejo de errores.
   - **d.** Documentación de una función.

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] J.J. Allaire et al. Quarto. DOI: 10.5281/ZENODO.5960048. Zenodo, Jan. 10, 2022. doi: 10.5281/ZENODO.5960048. url: https://zenodo.org/record/5960048.

[2] John M. Chambers. Software for data analysis: programming with R. Statistics and computing. OCLC: ocn191243189. New York: Springer, 2008.

[3] chocolatey. Chocolatey Software Docs | Chocolatey - Software Management for Windows. 2022. url: https://docs.chocolateyorg/en-us.

[4] Colin Gillespie and Robin Lovelace. Efficient R programming: a practical guide to smarter programming. First edition. OCLC: ocn964820738. Sebastopol, CA: O'Reilly Media, Inc, 2016.

[5] Norman S. Matloff. The art of R programming: tour of statistical software design. OCLC: ocn711045702. San Francisco: No Starch Press, 2011.

[6] Hadley Wickham. "Tidy Data". In: Journal of Statistical Software 59 (Sept. 12, 2014), pp. 1–23. doi: 10.18637/jss.v059.i10. url: https://doi.org/10.18637/jss.v059.i10.

[7] Hadley Wickham and Garrett Grolemund. R for data science: import, tidy, transform, visualize, and

CD303 Programación Avanzada en R

Descripción

En este curso se presentan desde los elementos básicos hasta la elaboración de un proyecto del proceso de manipulación, visualización de datos y la elaboración de reportes para resolver un problema de Ciencia de Datos en R.

Objetivos

En este curso el estudiante será capaz de:

1. Verificar el formato correcto de los datos y realizar correcciones.
2. Crear resúmenes de datos e interpretarlos.
3. Comprender los conceptos básicos de la visualización de datos.
4. Aplicar las mejores prácticas para la elaboración de gráficos y reportes.
5. Diseñar procesos de análisis de datos replicables.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.
3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Implementar proyectos de Minería de Datos que involucre los pasos básicos de manipulación, visualización y reportes de nivel empresarial utilizando el lenguaje de programación R.

Contenido

1. **Creación de proyectos.**
   - **a.** Control de rutas.
   - **b.** Estructura de carpetas.
   - **c.** Buenas prácticas para el diseño de proyectos.
2. **Diseño de reportes con Quarto y KnitR.**
   - **a.** Sintaxis de Quarto y Knit.
   - **b.** Buenas prácticas para el diseño de reportes.
   - **c.** Modificar tipos de datos de un archivo.
3. **Manipulación de datos con Tidyverse.**
   - **a.** Selección de variables.
   - **b.** Selección de individuos.
   - **c.** Resumenes de datos.
   - **d.** Ordenar tablas.
   - **e.** Creación de nuevas variables.
4. **Programación de Métodos Predictivos.**
   - **a.** KNN.
   - **b.** Árboles.
   - **c.** Bosques Aleatorios.
   - **d.** Potenciación y XGBoosting.
5. **Calibración y Selección de Modelos Predictivos desde la consola de R.**
   - **a.** Calibración y Selección de Modelos de Clasificación.
   - **b.** Calibración y Selección de Modelos de Regresión.

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] J.J. Allaire et al. Quarto. DOI: 10.5281/ZENODO.5960048. Zenodo, Jan. 10, 2022. doi: 10.5281/ZENODO.5960048. url: https://zenodo.org/record/5960048.

[2] John M. Chambers. Software for data analysis: programming with R. Statistics and computing. OCLC: ocn191243189. New York: Springer, 2008.

[3] chocolatey. Chocolatey Software Docs | Chocolatey - Software Management for Windows. 2022. url: https://docs.chocolatey.org/en-us.

[4] Colin Gillespie and Robin Lovelace. Efficient R programming: a practical guide to smarter programming. First edition. OCLC: ocn964820738. Sebastopol, CA: O'Reilly Media, Inc, 2016.

[5] Norman S. Matloff. The art of R programming: tour of statistical software design. OCLC: ocn711045702. San Francisco: No Starch Press, 2011.

[6] Hadley Wickham. "Tidy Data". In: Journal of Statistical Software 59 (Sept. 12, 2014), pp. 1–23. doi: 10.18637/jss.v059.i10. url: https://doi.org/10.18637/jss.v059.i10.

[7] Hadley Wickham and Garrett Grolemund. R for data science: import, tidy, transform, visualize, and model data. First edition. OCLC: ocn968213225. Sebastopol, CA: O'Reilly, 2016.

CD306
Manipulación y Preparación de Datos

Descripción

En este curso se presentarán los fundamentos del lenguaje R para el procesamiento de datos. El énfasis principal del curso será examinar diversos componentes del lenguaje, como lo son funciones, expresiones, librerías, entre otros. Se le dará especial importancia al uso del lenguaje como herramienta de manipulación de información, como punto de partida para el desarrollo de aplicaciones de minería de datos. Para esto se utilizarán diversos paquetes en R para manipulación de datos, así como motores de bases de datos como SQLite y MySQL.

Objetivos

En este curso el estudiante será capaz de:

1. Utilizar el lenguaje R como mecanismo de extracción de datos e información a partir de repositorios con grandes volúmenes de datos.
2. Hacer uso correcto del lenguaje para construir consultas complejas que permitan manipular información de distintas tablas de datos simultáneamente.
3. Entender el lenguaje R desde el punto de vista de teoría de conjuntos y lógica de predicados, permitiendo realizar operaciones usuales como lo son uniones, intersecciones, diferencias, entre otros.
4. Utilizar SQLite y MySQL como motores de bases de datos basados en SQL.
5. Importar información de un Administrador de base de datos SQL a R.
6. Exportar resultados obtenidos en R a un Administrador de base de datos SQL.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.
3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Minería de Datos que involucren alta manipulación de datos utilizando el lenguaje R.

Contenido

1. **Estructuración de datos.**
   - **a.** ¿Qué es Data Wrangling?
   - **b.** Uso de tidyr para limpieza, estructuración y manipulación de datos en R
     - Uniones.
     - Separaciones.
     - Transformaciones.
   - **c.** Uso de pipes (%>%) en R.
2. **Manipulación de datos.**
   - **a.** Uso de dplyr para el procesamiento y manipulación de datos en R.
   - **b.** Explorar tablas completas o subconjuntos (SELECT).
   - **c.** Derivación/Creación de nuevas columnas. (MUTATE).
   - **d.** Ordenamiento de datos (ARRANGE).
   - **e.** Filtros de datos. (FILTER).
   - **f.** Agrupación de individuos para obtener métricas (GROUP_BY, SUMMARISE).
3. **Manipulación de Textos.**
   - **a.** Manejo de variables categóricas.
   - **b.** Extraer, detectar y reemplazar componentes de un texto.

c. Realizar conteos de patrones en los textos.

d. Manejo de argumentos en textos.

4. **Conexión con Bases de Datos.**

a. Conexión a Bases de datos (DBI y RODBC).

b. Base de datos SQLite y MySQL.

c. Creación de estructuras (data frames) en R.

d. Almacenar datos localmente.

e. Paquete dbplyr (Uso de sintaxis dplyr a una base de datos).

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] Bradley C. Boehmke. Data Wrangling with R. Springer, 2016. url: https://link.springer.com.una.remotexs.co/book/10.1007/978-3-319-45599-0 (visited on 10/10/2022).

[2] Vikram Dayal. Quantitative Economics with R. Singapore: Springer Singapore, 2021. 326 pp. isbn: 9789811634345. doi: 10.1007/978-981-16-3434-5. url: https://link.springer.com/10.1007/978-981-16-3434-5 (visited on 10/10/2022).

[3] "MySQL: My Structured Query Language". In: Encyclopedia of Systems Biology. Ed. by Werner Dubitzky et al. New York, NY: Springer New York, 2013, pp. 1485–1486. isbn: 978-1-4419-9863-7. doi: 10.1007/978-1-4419-9863-7_100986. url: https://doi.org/10.1007/978-1-4419-9863-7_100986.

[4] Alfredo Ferro et al. "MySQL Data Mining: Extending MySQL to Support Data Mining Primitives (Demo)". In: Knowledge-Based and Intelligent Information and Engineering Systems. Ed. by Rossitza Setchi et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 438–444. isbn: 978-3-642-15393-8.

[5] Andreas Meier and Michael Kaufmann. SQL & NoSQL Databases: Models, Languages, Consistency Options and Architectures for Big Data Management. Wiesbaden: Springer Fachmedien Wiesbaden, 2019. 229 pp. isbn: 978-3-658-24548-1 978-3-658-24549-8. doi: 10.1007/978-3-658-24549-8. url: http://link.springer.com/10.1007/978-3-658-24549-8 (visited on 10/10/2022).

[6] Laurie A. Schintler and Connie L. McNeely, eds. Encyclopedia of Big Data. Springer Cham, 2022. 976 pp. url: https://link.springer.com.una.remotexs.co/book/10.1007/978-3-319-32010-6 (visited on 10/10/2022).

[7] Geoffrey I. Webb and Claude Sammut, eds. Encyclopedia of Machine Learning and Data Mining. Second Edition. Springer New York, 2017. 1335 pp. url: https://link.springer.com.una.remotexs.co/book/10.1007/978-1-4899-7687-1 (visited on 10/10/2022).

CD309 Visualización e Interpretación de Datos

Descripción

En este curso se presentarán desde los elementos básicos del proceso de visualización de datos hasta los elementos más avanzados incluyendo la teoría para seleccionar el tipo y los elementos que más se adaptan al problema. También se aprenderá a generar gráficos interactivos y la elaboración de presentaciones usando el lenguaje R.

Objetivos

En este curso el estudiante será capaz de:

1. Verificar el formato correcto de los datos y realizar correcciones.
2. Crear resúmenes de datos e interpretarlos.
3. Comprender los conceptos básicos de la visualización de datos.
4. Aplicar las mejores prácticas para la elaboración de gráficos y reportes.
5. Diseñar procesos de análisis de datos replicables.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.
3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar reportes de nivel empresarial utilizando el lenguaje de programación R.

Contenido

1. **Principios de visualización de datos.**
   - **a.** Teoría de la gramática de los gráficos.
   - **b.** Selección de gráficos.
   - **c.** Diseño de gráficos y accesibilidad.
   - **d.** Teoría del color aplicada a la visualización de datos.
2. **Visualización de datos geoespaciales.**
   - **a.** Tipo de datos geoespeciales.
   - **b.** Sistemas de coordenadas geográficas.
   - **c.** Creación de mapas.
3. **Diseño de gráficos interactivos.**
   - **a.** Herramientas para gráficos interactivos.
   - **b.** Gráficos interactivos.
   - **c.** Mapas interactivos.
   - **d.** Presentación de documentos interactivos y sus limitacines.
4. **Diseño de presentaciones.**
   - **a.** Estructura de una presentacíon.
   - **b.** Herramientas para la creación de documentos y presentaciones con quarto.
   - **c.** Buenas practicas para la elaboración de documentos.
5. **“Storytelling” con Datos.**
   - **a.** El arte de contar historias con datos.
   - **b.** Visualizar datos es más que graficar datos.
   - **c.** Buenas practicas para la elaboración de gráficos y reportes.

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] Bradley C. Boehmke. Data Wrangling with R. Springer, 2016. url: https://link.springer.com.una.remotexs.co/book/10.1007/978-3-319-45599-0 (visited on 10/10/2022).

[2] Vikram Dayal. Quantitative Economics with R. Singapore: Springer Singapore, 2021.326 pp. isbn: 9789811634345. doi: 10.1007/978-981-16-3434-5. url: https://link.springer.com/10.1007/978-981-16-3434-5 (visited on 10/10/2022).

[3] "MySQL: My Structured Query Language". In: Encyclopedia of Systems Biology. Ed. by Werner Dubitzky et al. New York, NY: Springer New York, 2013, pp. 1485–1486. isbn:978-1-4419-9863-7. doi: 10.1007/978-1-4419-9863-7_100986. url: https://doi.org/10.1007/978-1-4419-9863-7_100986.

[4] Alfredo Ferro et al. "MySQL Data Mining: Extending MySQL to Support Data Mining Primitives (Demo)". In: Knowledge-Based and Intelligent Information and Engineering Systems. Ed. by Rossitza Setchi et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 438–444. isbn: 978-3-642-15393-8.

[5] Andreas Meier and Michael Kaufmann. SQL & NoSQL Databases: Models, Languages, Consistency Options and Architectures for Big Data Management. Wiesbaden: Springer Fachmedien Wiesbaden, 2019. 229 pp. isbn: 978-3-658-24548-1 978-3-658-24549-8. doi: 10.1007/978-3-658-24549-8. url: http://link.springer.com/10.1007/978-3-658-24549-8 (visited on 10/10/2022).

[6] Laurie A. Schintler and Connie L. McNeely, eds. Encyclopedia of Big Data. Springer Cham, 2022. 976 pp. url: https://link.springer com.una.remotexs.co/book/10.1007/978-3-319-32010-6 (visited on 10/10/2022).

[7] Geoffrey I. Webb and Claude Sammut, eds. Encyclopedia of Machine Learning and Data Mining. Second Edition. Springer New York, 2017. 1335 pp. url: https://link.springer.com.una.remotexs.co/book/10.1007/978-1-4899-7687-1 (visited on 10/10/2022).

CD400 Implementación de Proyectos en Ciencia de Datos

Descripción

En este curso se estudiarán en detalle las técnicas de Validación Cruzada (Cross-Validation) y Remuestreo (boostrapping) con el objetivo de calibrar y seleccionar el mejor método de Minería de Datos para un problema y juego de datos dado. Se estudiará como programar y automatizar la Validación Cruzada (Cross-Validation) y Remuestreo (boostrapping) en el lenguaje R. Los ejemplos de este curso estarán motivados por problemas reales en el campo. Por lo tanto, los estudiantes adquieren conocimientos de muchas herramientas diferentes que pueden combinarse para resolver problemas reales.

Objetivos

En este curso el estudiante será capaz de:

1. Implementar programas en R para Validación Cruzada (Cross-Validation) y Remuestreo (bootsrapping) tanto en Clasificación como en Regresión.
2. Aplicar adecuadamente una Validación Cruzada y un Remuestreo.
3. Calibrar adecuadamente tanto métodos exploratorios como métodos predictivos en Ciencia de Datos.
4. Seleccionar el mejor modelo predictivo dado un conjunto de datos.
5. Implementar proyectos avanzados en Ciencia de Datos usando R.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.
2. Trabajos prácticos semanales.

3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Minería de Datos que involucren métodos exploratorios y predictivos avanzados usando código en R.

Contenido

1. **Implementación de métodos Métodos Exploratorios (Clustering o Aprendizaje no Supervisado).**
   - **a.** Análisis en Componentes Principales.
   - **b.** Clustering Jerárquico.
   - **c.** El método de K-medias.
2. **Implementación de Métodos de Predictivos y Métodos de Regresión.**
   - **a.** El método de los K vecinos más cercanos.
   - **b.** Método de Bayes.
   - **c.** Análisis Discriminante Lineal y Cuadrático.
   - **d.** Máquinas Vectoriales de Soporte.
   - **e.** Árboles de Decisión.
   - **f.** Bosques Aleatorios (Random Forest).
   - **g.** Métodos de Potenciación (Boosting).
   - **h.** Redes Neuronales.
   - **i.** Guardando y Recuperando en disco un Modelo para su posterior uso.
   - **j.** Implementación de todos los métodos anteriores para Regresión.
   - **k.** Implementación de la Regresión Clásica.
3. **Implementación de la Validación Cruzada (cross-validation) para regresión y para clasificación, Curvas ROC.**
   - **a.** Enfoque: "tabla de aprendizaje y tabla de testing" (the validation test approach).
   - **b.** Validación cruzada dejando uno fuera (Leave-one-out cross-validation - LOOCV).
   - **c.** Validación cruzada usando K grupos (K-fold cross-validation).
   - **d.** Versiones en paralelo.

e. Implementación de la Calibración y Selección de Modelos.

f. Implementación de las Curvas ROC.

4. **Implementación de Métodos en Series de Tiempo.**

a. Trabajando con fechas.

b. Implementación de los métodos de Holt-Winters.

c. Implementación de los Métodos ARIMA.

d. Evaluación del error.

e. Seleccionando el mejor método.

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] Esteban Alfaro, Matias Gamez, and Noelia García. "adabag: An R Package for Classification with Boosting and Bagging". In: Journal of Statistical Software 54 (Sept. 3, 2013), pp. 1–35. issn: 1548-7660. doi: 10.18637/jss.v054.i02. url: https://doi.org/10.18637/jss.v054.i02 (visited on 10/19/2022).

[2] Alexandros Karatzoglou, David Meyer, and Kurt Hornik. "Support Vector Machines in R". In: Journal of Statistical Software 15 (Apr. 6, 2006), pp. 1–28. issn: 1548-7660. doi: 10.18637/jss.v015.i09. url: https://doi.org/10.18637/jss.v015.i09.

[3] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: 2 (2002), p. 5.

[4] Wei-Yin Loh. "Classification and regression trees". In: WIREs Data Mining and Knowledge Discovery 1.1 (2011). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.8, pp. 14–23. issn: 1942-4795. doi: 10.1002/widm.8. url: https://onlinelibrary.wiley.com/doi/abs/10.1002/widm.8.

[5] Boris Mirkin and Boris Mirkin. Clustering for Data Mining: A Data Recovery Approach. New York: Chapman and Hall/CRC, Apr. 28, 2005. 296 pp. isbn: 978-0-429-13754-9. doi: 10.1201/9781420034912.

[6] Simon Sheather. A Modern Approach to Regression with R. Google-Books-ID: IZ6Ra0efSdsC. Springer Science & Business Media, Feb. 27, 2009. 398 pp. isbn: 978-0-387-09608-7.

[7] The 'K' in K-fold Cross Validation. url: https://arpi.unipi.it/handle/11568/962587 (visited on 10/19/2022).

[8] Tzu-Tsung Wong. "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation". In: Pattern Recognition 48.9 (Sept. 1, 2015), pp. 2839–2846. issn: 0031-3203. doi: 10.1016/j.patcog.2015.03.009. url: https://www.sciencedirect.com/science/article/pii/S0031320315000989 (visited on 10/19/2022). [9] Zhongheng Zhang. "Naïve Bayes classification in R". In: Annals of Translational Medicine 4.12 (June 2016), p. 241. issn: 2305-5839. doi: 10.21037/atm.2016.03.38. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4930525/ (visited on 10/19/2022).

CD900 Proyecto Final de Graduación

Descripción

En este curso el estudiante desarrollará un proyecto de principio a fin, siguiendo las buenas prácticas de gestión de proyectos en Ciencia de Datos según la metodología CRISP-DM. Para el proyecto el estudiante utilizará datos de la empresa en donde trabaja o datos en los cuales está particularmente interesado(a).

Objetivos

En este curso el estudiante será capaz de:

1. Utilizar adecuadamente la metodología CRISP-DM para el desarrollo de proyectos en Ciencia de Datos.
2. Distinguir entre un proyecto de segmentación, clasificación, regresión y series de tiempo.
3. Elegir, para un problema concreto, qué técnicas de Ciencia de Datos son las más apropiadas.
4. Determinar la herramienta de software más adecuada para el enfrentar el problema planteado.
5. Generar los modelos y patrones elegidos utilizando una herramienta o paquete de Ciencia de Datos.
6. Evaluar la calidad de un modelo, utilizando técnicas sencillas de evaluación, validación cruzada o remuestreo.

Metodología

Basado en la teoría y en la aplicación directa de los conceptos aprendidos. Para esto se dispondrán de las siguientes herramientas:

1. Una vídeo conferencia semanal, las cuales quedarán grabadas en Zoom, para que los alumnos la puedan acceder en cualquier momento.

2. Trabajos prácticos semanales.
3. Foros para plantear dudas al tutor y compañeros.
4. Aula virtual en Moodle.

Luego de este curso el estudiante será capaz de:

Desarrollar proyectos de Ciencia de Datos de principio a fin, es decir, iniciando con la limpieza de datos hasta la generación de conocimiento derivados de los datos de la organización.

Contenido

1. Fundamentos de la metodología para la implementación de proyectos en Ciencia de Datos CRISP-DM.
2. Estudio comparativo de diferentes tipos problemas en Ciencia de Datos.
3. Presentación del anteproyecto.
4. Presentación y defensa del proyecto final.

Evaluación

El curso se evalúa con 4 tareas, una por semana, cada tarea tiene un valor de 25 puntos. La nota mínima de aprobación es de 70.

Bibliografía

[1] José Alberto Gallardo Arancibia. "Metodología para la definición de requisitos en proyectos de data mining (er-dm)". http://purl.org/dc/dcmitype/Text. Universidad Politécnica de Madrid, 2009. url: https://dialnet.unirioja.es/servlet/tesis?codigo=20961 (visited on 10/19/2022).

[2] Ana Azevedo and Manuel Filipe Santos. "KDD, SEMMA and CRISP-DM: a parallel overview". In: IADS - DM (2008). Accepted: 2012-06-14T09:51:14Z. url: https://recipp.ipp.pt/handle/10400.22/136 (visited on 10/19/2022).

[3] CRISP-DM 1.0: Step-by-step Data Mining Guide. Google-Books-ID: po7FtgAACAAJ.SPSS, 2000. book.

[4] Graham J. Williams and Simeon J. Simoff. Data Mining: Theory, Methodology, Techniques, and Applications. Google-Books-ID: 44z0BwAAQBAJ. Springer, Jan. 22, 2006.341 pp. isbn: 978-3-540-32548-2.

Correo electrónico : info@promidat.com

Teléfono: +506 4030-1205

Web: www.promidat.com

WhatsApp: +506 8712-6978

Directo: +506 2271-0464

Programa Iberoamericano de
Formación en Minería de Datos