

Práctica 1.

Intro a RapidMiner: Clasificación

*Máster en Big Data*

*Curso 2019-2020*

***Data Science & Advanced Analytics***

*Montse Llos i Bombardó*

*RapidMiner*

RapidMiner es una plataforma de desarrollo de proyectos de Machine Learning donde

data analysts, data scientists e ingenieros pueden explorar, analizar, construir y

preparar sus productos de forma más eficiente, con una reducción de tiempo de desarrollo muy considerable.

Dirigida principalmente al ámbito de los citizen data scientists. Un citizen data scientist es un perfil más generalista, un ingeniero, atento a los procesos críticos de la empresa, que con un mínimo conocimiento de Machine Learning, sabe traducir la problemática en acciones que ayuden a la toma de decisiones. Se trata tanto de hacer las preguntas correctas como de crear el algoritmo correcto. Comprender cómo resolver los verdaderos problemas de negocio frente a la fórmula matemática es importante, pero la capacidad de evaluar las relaciones entre múltiples variables y utilizar los datos para dar forma a las perspectivas predictivas es esencial.

Dispone de cantidad de herramientas para la transformación y análisis de datos,

permite integrar scripts en diferentes lenguajes de programación y ofrece además la

posibilidad de aplicar algoritmos de Machine Learning de forma relativamente fácil.

Otro de los puntos más importantes es que se integra perfectamente con sistemas Big

Data como Hadoop, Spark, Hive y demás, aprovechando las herramientas de RapidMiner pero dejando el procesamiento de datos a los sistemas Big Data.

RapidMiner nació en una universidad de Dortmund y lleva seis años consecutivos como líder en el sector según Gartner Group.

*Descarga de RapidMiner*

Para descargarlo, nos dirigimos a su web oficial <https://rapidminer.com/get-started/>

Para obtener una licencia gratuita de un año, es importante que nos registremos con nuestra cuenta de email del campus EAE (@campus.eae.es). Así obtendremos una licencia completa.

En Virtual Labs, también está instalado.

Existe una versión Cloud en <https://automodel.rapidminer.com/login>, sin necesidad de instalar nada pero que actualmente únicamente se puede ejecutar en modo automodel (generación automática del mejor modelo según el tipo de predicción y datos que se disponen)

*Práctica 1.1*

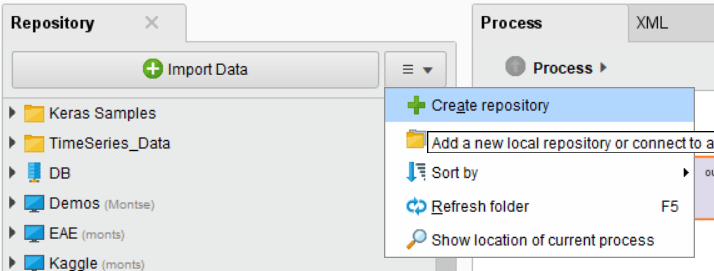
La empresa de e-commerce VamosBien quiere conocer con antelación qué clientes tienen una alta probabilidad de que abandonen su plataforma. Esto les ayudaría a redirigir su esfuerzo comercial hacia los clientes más susceptibles de abandonarlos.

Nos han dado un Excel con datos sociodemográficos de sus de clientes y forma de pago, fecha de su última transacción con VamosBien y fecha de los clientes que ya se han dado de baja.

Con estos datos, deberíamos ofrecerles un modelo que pueda predecir para los clientes activos y futuros clientes la probabilidad de abandono de su negocio

En el Apartado de Recursos del Campus encontraremos los datos y procesos necesarios para el desarrollo de la práctica en RapidMiner. En Carpeta Prácticas\Práctica 1

Primero de todo crearemos un Repositorio dentro de RapidMiner que le llamaremos Prácticas EAE, por ejemplo.



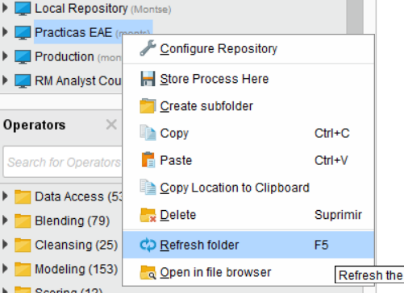
Después, iremos al Explorador y situaremos los procesos de RapidMiner de esta Práctica 1 en una carpeta nueva que cuelgue de [User Path]\.RapidMiner\Repositories\Practicas EAE.

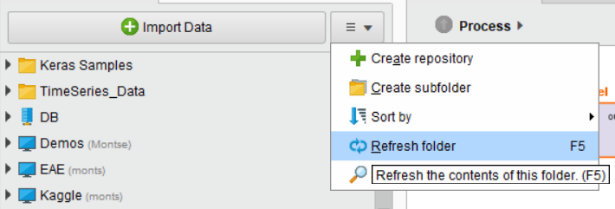
También se puede situar el Repositorio en una carpeta que no sea la de por defecto.

Los datos Excel no es necesario que estén en esta carpeta.

Finalmente, dentro de RapidMiner, sobre el Repositorio Prácticas EAE que hemos creado haremos un Refresh Folder mediante uno de los tres puntos mencionados a continuación:

* Seleccionando el repositorio + F5
* Seleccionando el repositorio + botón derecho mouse + Refresh Folder



* Seleccionando el repositorio + desplegando menú situado en la parte superior de los repositorios

A continuación, seguiremos paso a paso cada uno de los apartados de la Práctica 1.1.

Para bajarse la práctica 1.1 debemos realizar los siguientes pasos:

1. Abrir el explorador e ir a la carpeta donde está el Repositorio Prácticas EAE
2. Crea un archivo comprimido de la carpeta Práctica 1.1.

*Ejercicio 1.2*

Cargaremos los datos de Titanic, un dataset tradicionalmente usado en el mundo del Machine Learning.

A partir de los datos de los pasajeros sobre clase en la que iba, puerto de embarque, edad, sexo y otros, crearemos un modelo que pueda aprender y predecir las características para determinar si un pasajero sobrevivirá o no.

Para llevarlo a cabo, tendréis que seguir paso a paso cada uno de los subapartados que se encuentran en la subcarpeta Ejercicio 1.2.

El ejercicio se debe realizar en RapidMiner.

Número límite de personas: 1

Entregables:

* Un fichero .zip que contenga la carpeta del repositorio de RapidMiner con todos los ejercicios realizados.
  + Nombre del fichero: <Apellido+InicialNombre>\_Ejercicio\_1.2.zip
  + Ejemplo: *LlosM\_Ejercicio\_1.2.zip*
* Un documento Pdf con nombre y foto de la persona que ha realizado el ejercicio.
  + Nombre del fichero: <Apellido+InicialNombre>\_Ejercicio\_1.2.pdf
  + Ejemplo: *LlosM\_Ejercicio\_1.2.pdf*

Fecha límite entrega: 27 enero 2020 a las 24h

Enviar a: subir al campus