

Práctica 5.

Intro a Azure ML: Clustering

*Máster en Big Data*

*Curso 2019-2020*

***Data Science & Advanced Analytics***

*Montse Llos i Bombardó*

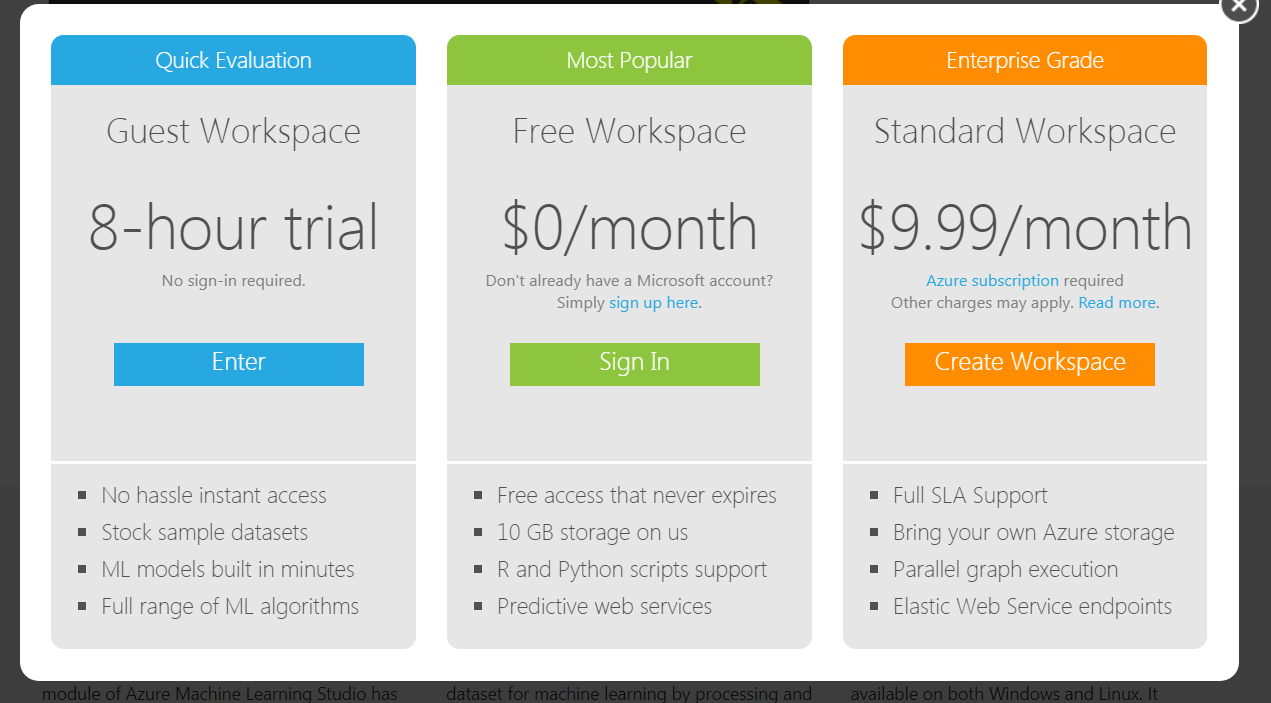
*Microsoft Azure Machine Learning Studio*

Microsoft Azure Machine Learning Studio es la plataforma de Microsoft para el desarrollo de proyectos de Machine Learning basado en un sistema drag&drop.

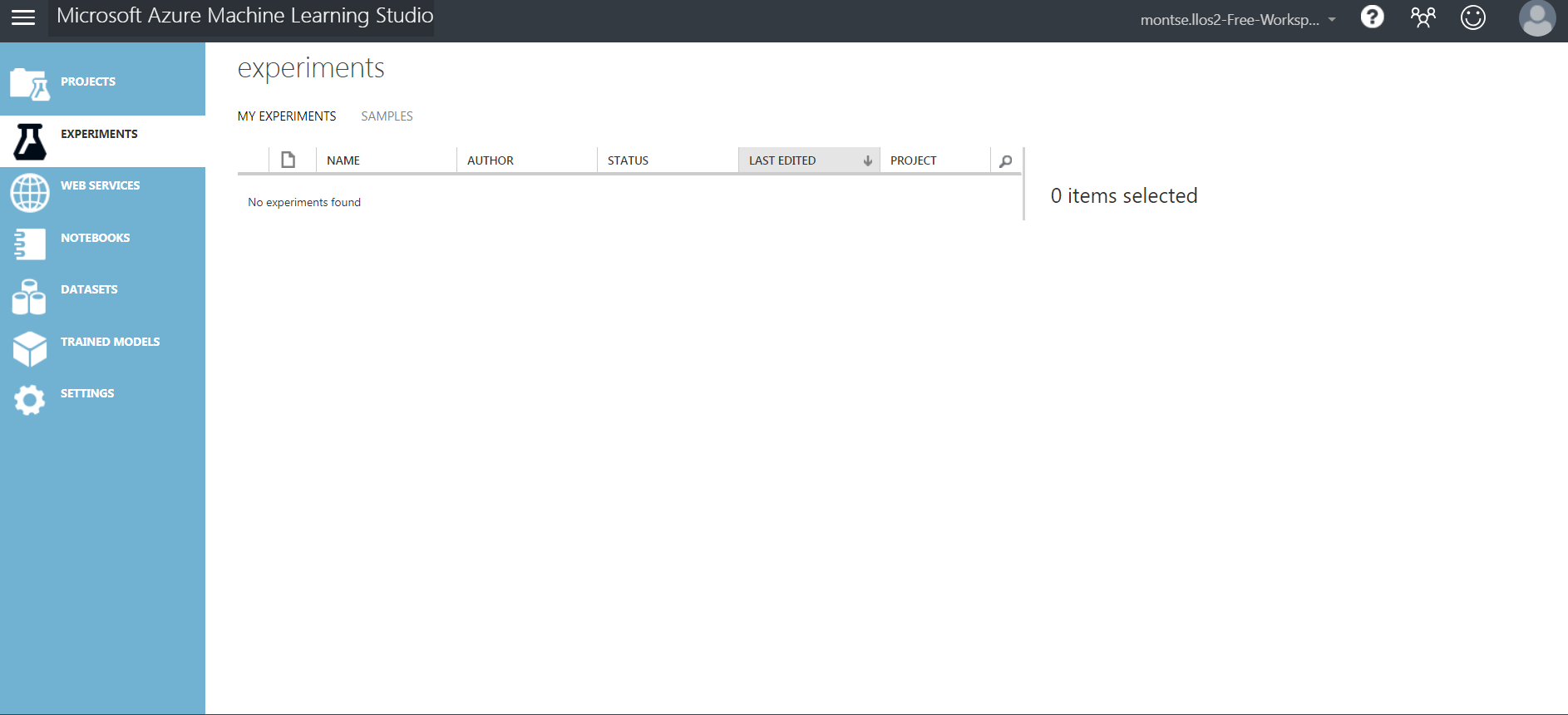
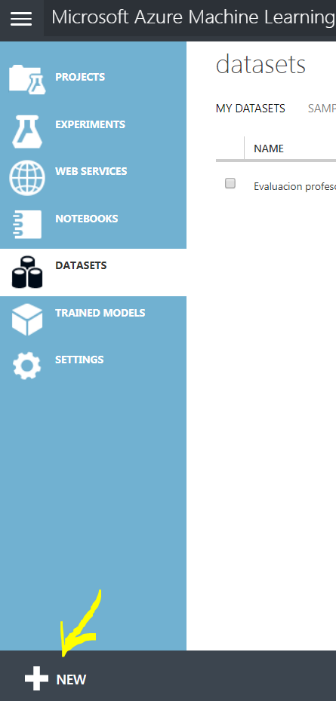
Para acceder al entorno, no nos instalaremos nada ya que el acceso es vía cloud.

Debemos acceder al siguiente enlace: <https://studio.azureml.net/?selectAccess=true&o=1>

El acceso es free, con 10GB máximo de almacenamiento



El entorno de trabajo es el siguiente:



Para cargar datos, ir a la sección Dataset y click en el ‘+’

<https://docs.microsoft.com/es-es/azure/machine-learning/studio/import-data>

Crear proyecto y experimento para empezar un nuevo proceso

Ejemplos: <https://gallery.azure.ai/experiments>

En experiments/samples hay varios ejemplos a cargar

Tutorial interesante: <https://www.youtube.com/watch?v=F_o9GYBDTF4>

Para bajarse datos hay un operador ‘Convert to CSV’+Run+botón derecho del mouse sobre el operador y Download

Otros: existe una licencia gratuita de Azure Microsoft por un año dándose de alta con el login del campus EAE <https://azure.microsoft.com/en-us/free/students/>

*Práctica 5.0*

Vamos a cargar el dataset de Titanic usado en el Ejercicio 1 de Clasificación y lo resolveremos usando Azure.

El fichero lo tendremos que cargar en formato CSV y separado por comas. Lo encontraremos en la carpeta de la práctica 4.

Los pasos a realizar son los siguientes:

## Paso 1: Upload

1. Cargar datos

## Paso 2: Titanic ETL

1. Eliminate ‘Cabin’, ‘Life Boat’, ‘Port of Embarkation’ and ‘Ticket Number’ columns from the result set
2. Eliminate the rows of data where the ‘Passenger Fare’ is missing
3. Replace the missing values of the ‘Age’ column with the average value of Age
4. Designate ‘Name’ as the Id column and ‘Survived’ as the Label column

## Paso 3: Titanic k-NN

1. Embed the Titanic ETL Result from Lab 2
2. Split the Data (70% Train and 30% Test)
3. Perform k-NN on the Training data
4. Apply the Model to the Test Data
5. Review the Results
6. Alter k value and review the Results

## Paso 4: Titanic Validation

1. Embed the Titanic ETL Result from Lab 2
2. Perform X-Validation with k-NN
3. Apply the Model
4. Evaluate the Performance
5. Review the Results
6. Alter k value and review the Results

## Paso 5: Titanic Normalization

1. Embed the Titanic ETL Result from Lab 2
2. Perform X-Validation with k-nn
3. Normalize before k-nn
4. Group the Models
5. Apply the Model
6. Evaluate the Performance
7. Review the Results
8. Alter k value and review the Results

## Paso 6: Titanic Optimize Parameter

1. Embed the Titanic ETL Result from Lab 2
2. Perform X-Validation with Normalize
3. Group the Models
4. Apply the Model
5. Evaluate the Performance
6. Introduce an Optimize Parameter operator and move the X-Validation as a sub-process inside the Optimizer
7. Optimize the k-value parameter, Min-1, Max-30 & Steps-30 (linear)
8. Add the Log operator to capture the results
9. Run the optimization and review the results

## Paso 7: Titanic Naive Bayes

1. Duplicate the Lab 5 results in Lab 7 folder by using “Store Process Here”
2. Replace k-NN with ‘Naive Bayes’
3. Review the Results

## Paso 8: Titanic Linear Regression

1. Embed the ETL Solution from Lab 2
2. Perform X-Validation
3. In the X-Val operator, convert the “Survived” column to a Binominal
4. Convert the remaining Nominal columns (except Name) to Numeric
5. Perform Linear Regression
6. Group the models in order
7. Apply Model and evaluate Performance
8. Review the Results

## Paso 9: Titanic Decision Tree

1. Duplicate the Lab 5 results in Lab 7 folder by using “Store Process Here”
2. Replace k-NN with ‘Deciosion Tree'
3. Review the Results

Intentaremos realizar los pasos que podamos y así también veremos las diferencias que existen entre las dos herramientas.

*Práctica 5.1*

En EAE se han realizado unas encuestas para conocer las opiniones que tienen sus alumnos de las asignaturas y profesores.

Las medias de las encuestas ya ofrecen información sobre sus opiniones, pero desvirtúan las distintas opiniones que haya.

Es por ello que nos piden hacer una clusterización para analizar con más detalle las características más relevantes de cada grupo

Para ello, nos han dado unos datos con los siguientes campos:

* instr:
  + Identificador del profesor. Valores posibles {1,2,3}
* class:
  + Código de curso. Valores posibles entre {1-13}
* repeat:
  + Número de veces que el estudiante ha repetido el curso. Valores entre {0,1,2,3…}
* attendance:
  + Código con el nivel de asistencia. Valores posibles {0,1,2,3,4}
* difficulty:
  + Nivel de dificultad del curso tal y como lo ha percibido el estudiante. Valores posibles {1,2,3,4,5}
* Q1:
  + El contenido del curso semestral, el método de enseñanza y el sistema de evaluación fueron proporcionados al principio
* Q2:
  + Las metas y objetivos del curso se establecieron claramente al principio del período.
* Q3:
  + El curso valió la cantidad de crédito que se le asignó.
* Q4:
  + El curso se impartió de acuerdo con el programa de estudios anunciado el primer día de clase.
* Q5:
  + Las discusiones en clase, las tareas, las solicitudes y los estudios fueron satisfactorios.
* Q6:
  + El libro de texto y otros recursos de los cursos fueron suficientes y actualizados.
* Q7:
  + El curso permitió trabajo de campo, aplicaciones, laboratorio, discusión y otros estudios.
* Q8:
  + El curso permitió trabajo de campo, aplicaciones, laboratorio, discusión y otros estudios.
* Q9:
  + Disfruté mucho de la clase y estaba ansioso por participar activamente durante las conferencias.
* Q10:
  + Mis expectativas iniciales sobre el curso se cumplieron al final del período o año.
* Q11:
  + El curso fue relevante y beneficioso para mi desarrollo profesional.
* Q12:
  + El curso me ayudó a ver la vida y el mundo con una nueva perspectiva.
* Q13:
  + El conocimiento del Instructor fue relevante y actualizado.
* Q14:
  + El instructor vino preparado para las clases.
* Q15:
  + El instructor enseñó de acuerdo con el plan de lecciones anunciado.
* Q16:
  + El instructor estaba comprometido con el curso y era comprensible.
* Q17:
  + El instructor llegó a tiempo para las clases.
* Q18:
  + El instructor tiene un discurso de entrega suave y fácil de seguir.
* Q19:
  + El instructor hizo uso efectivo de las horas de clase.
* Q20:
  + El instructor explicó el curso y estaba ansioso por ayudar a los estudiantes.
* Q21:
  + El instructor demostró un enfoque positivo hacia los estudiantes.
* Q22:
  + El instructor fue abierto y respetuoso de los puntos de vista de los estudiantes sobre el curso.
* Q23:
  + El instructor alentó la participación en el curso.
* Q24:
  + o El instructor dio tareas/proyectos relevantes y ayudó/guió a los estudiantes.
* Q25:
  + o El instructor dio tareas/proyectos relevantes y ayudó/guió a los estudiantes.
* Q26:
  + El sistema de evaluación del instructor (preguntas de medio término y finales, proyectos, tareas, etc.) midió efectivamente los objetivos del curso.
* Q27:
  + o El instructor proporcionó soluciones a los exámenes y las discutió con los estudiantes.
* Q28:
  + El instructor trató a todos los estudiantes de una manera correcta y objetiva

*Ejercicio 5.2*

*ClinicalProteomic Tumor AnalysisConsortium (NCI / NIH)* es un consorcio que sube a la red datos relacionados con sus estudios. En este caso, tenemos unos 12.000 registros de comportamientos distintos de 77 muestras cánceres de mama donde se ha practicado en laboratorio una mutación de ADN.

A partir de ello, se han medido una serie de parámetros llamados proteomas, que corresponden a la totalidad de proteínas de una célula.

Los datos se utilizaron para evaluar cómo las mutaciones en el ADN están afectando el panorama de expresión de proteínas en el cáncer de mama.

En un total de 86 atributos distintos, se muestra como el tumor ha cambiado. Destacar los siguientes atributos:

* RefSeq\_accession\_number: identificador
* gene\_symbol: símbolo único para cada gen (cada proteína se codifica por un gen)
* gene\_name: nombre completo del gen

El objetivo de esta práctica es clasificar las muestras con las parecidas para así facilitar el estudio de los diversos tipos de cáncer de mama.

Para llevar a cabo el ejercicio tendremos que:

* Bajarnos los datos del campo virtual
* Cargar los datos
* Hay nulos? Qué hacemos?
  + Nos comentan que los nulos son proteínas que no se han encontrado en la célula
* Hacer la clusterización
* Genera un proceso que clasifique las muestras sin tener que desarrollar de nuevo el modelo
* ¿Cómo podríamos describir las características más relevantes de cada uno de los clústeres?

Se puede desarrollar en RapidMiner, Azure Machine Learning Studio o Python. En todos los casos, comentar cada uno de los pasos realizados

Número límite de personas: 2

Entregables:

* Un fichero zip con el ejercicio realizado
  + Nombre del fichero: <Apellidos+InicalNombres>\_Ejercicio\_5.2.zip
  + Ejemplo si sois dos: *LLosM\_RooneyM\_Ejercicio\_5.2.zip*
* Un documento Pdf con nombre y foto de la persona que ha realizado el ejercicio y respuesta a las preguntas anteriores
  + Nombre del fichero: <Apellidos+Nombres>\_Ejercicio\_5.2.pdf
  + Ejemplo si sois dos: *LLosM\_RooneyM\_Ejercicio\_5.2.pdf*

Fecha límite entrega: 9 marzo 2020 a las 24h

Subir al campus