Resumen

Dentro de la ciencia de datos el aprendizaje automático tiene un papel destacado a la hora de detectar patrones de comportamiento o predicciones sobre un conjunto de datos inmensos que, sin la ayuda de este tipo de tecnología, sería sumamente complicado extraer cualquier tipo de conclusión. Como parte de la formación de futuros profesionales en el ámbito del Big Data, en este documento se dan las claves para comenzar a desarrollar estas habilidades mostrando cómo crear una competición de aprendizaje automático a través de la plataforma Kaggle. A través de la misma plataforma y, bajo el contexto de la propia competición, se abordarán los aspectos claves en el desarrollo de un proceso aprendizaje automático base.

práctica SESIÓN 1 - Introducción a la ciencia de datos

USO PRÁCTICO DE BIG DATA

Contenido

[1. Introducción 2](#_Toc70278748)

[2. Cómo crear una competición. 3](#_Toc70278749)

[2.1. Cómo crear los conjuntos de datos necesarios para la competición desde los datasets ya existentes en Kaggle. 7](#_Toc70278750)

[2.2. Lanzamiento de competición. 11](#_Toc70278751)

[3. Cómo participar en una competición 14](#_Toc70278752)

## Introducción



Kaggle es una plataforma para el aprendizaje, perfeccionamiento y competición de Ciencia de datos más importante en la actualidad. Cuenta con un millón de usuarios y está dirigida a perfiles especialmente relacionados con el análisis, interpretación y explotación de los datos.

Además de proporcionar diferentes recursos formativos, así como una propia plataforma sobre la que realizar los procesos de aprendizaje automático, Kaggle ofrece la posibilidad de participar en competiciones promovidas por diferentes organizaciones donde se busca medir la destreza sobre aprendizaje automático de los aspirantes. Estas competiciones no son solo interesantes por tener la posibilidad de obtener premios sino porque, en general, se trata de una comunidad solidaria y enfocada principalmente al aprendizaje común, dando la oportunidad de aprender nuevas técnicas y algoritmos. Estas competiciones también permiten participar en equipo con lo cual se favorece la colaboración y participación de diferentes usuarios en un mismo proyecto, siendo un plus para motivar a más participantes.

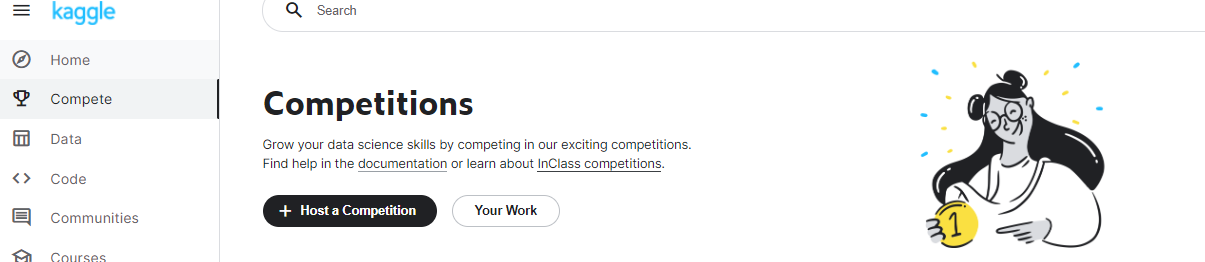
Entre los muchos servicios que ofrece Kaggle, destaca el entorno de programación que permite realizar desarrollo en Python, R o Julia en notebooks de Jupiter y sobre los propios servidores de Kaggle. Es completamente gratuito y resulta de gran ayuda ya que no tenemos que preocuparnos por configurar ni instalar nada, además de que permite copiar y compartir los kernels existentes entre los diferentes usuarios. Otra de las ventajas que ofrece es la gran cantidad de datasets de todo tipo que pueden ser explorados, además de darnos la opción de incorporar nuevos conjuntos de datos y compartirlos con el resto de usuarios. Existen foros en los que poder participar, diferentes cursos gratuitos para aprender nuevas técnicas, y un largo etcétera que convierten a Kagge en uno de los mejores lugares para el aprendizaje de la ciencia de datos.

En el ámbito de la educación Kaggle permite crear competiciones cerradas que cualquier docente puede utilizar como parte de su plan formativo siendo una excelente herramienta para fomentar el interés del alumnado a través de la competición.

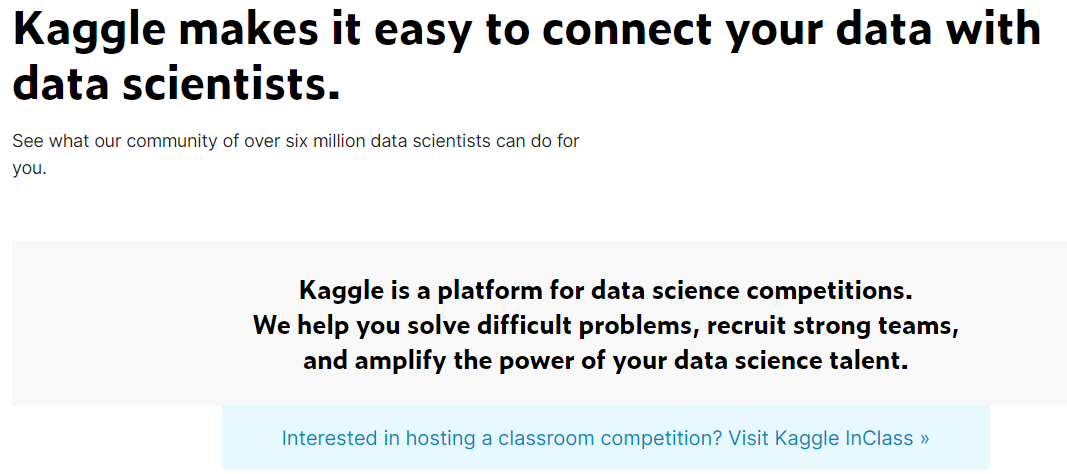
## Cómo crear una competición.

* Lo primero que tenemos que hacer acceder a la web de kaggle: <https://www.kaggle.com/> y registrarnos.
* Una vez hecho el registro en la página principal seleccionamos:

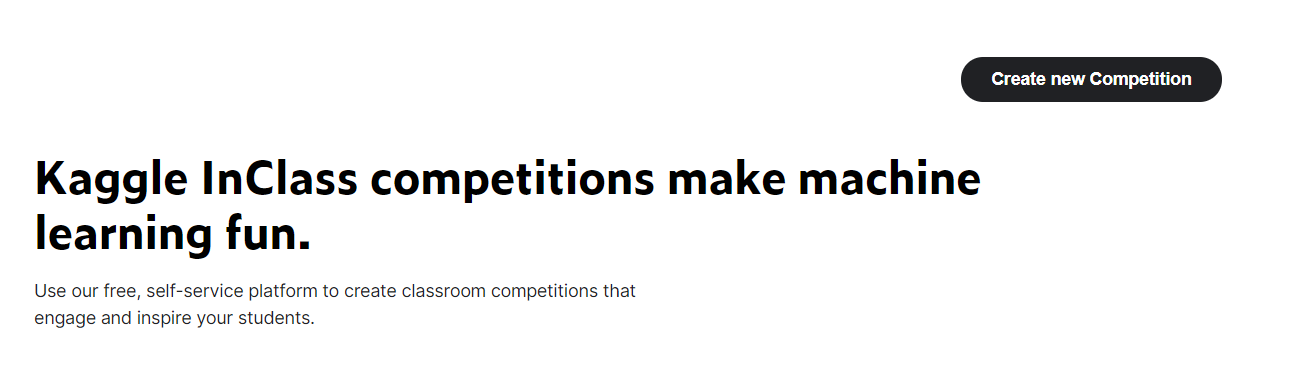
Compete 🡪Host a Competition



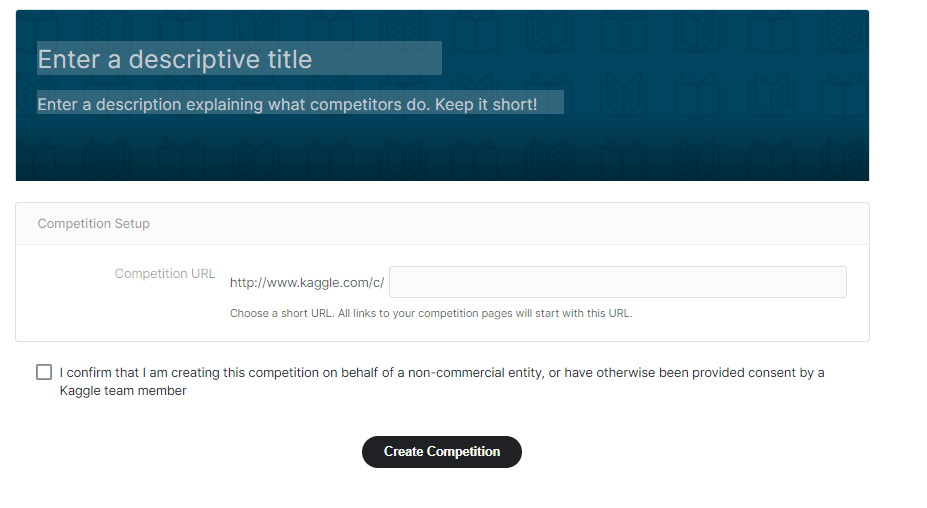
* Luego seleccionamos la opción donde indica si estamos interesados en organizar una competición en el aula.



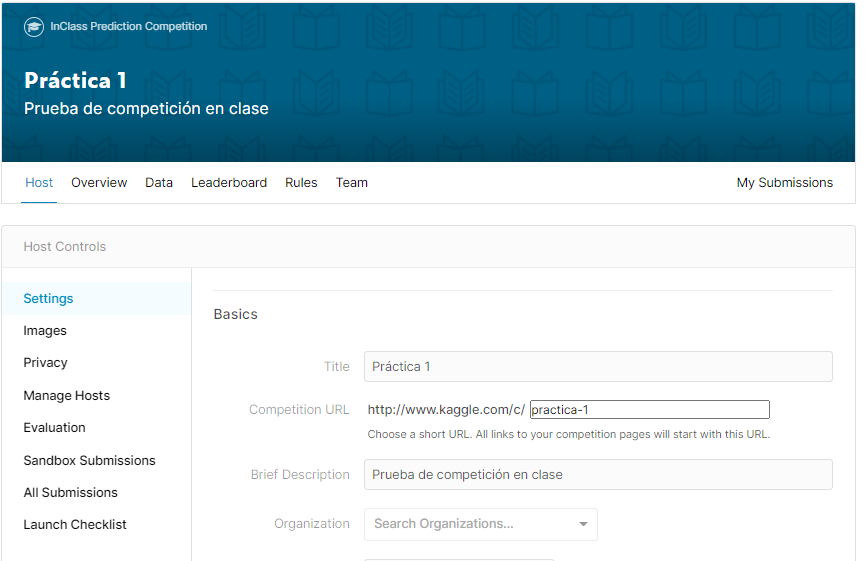
* El siguiente paso es simplemente seleccionar Crear Nueva Competición



* La acción anterior nos llevará a la página de nuestra competición, donde aportaremos la información que mejor creamos describe nuestra competencia. Para poder localizarla nos permite indicar una url para el acceso a la misma. Una vez se tiene listo, se da a crear competición.



* La siguiente ventana muestra la configuración principal.

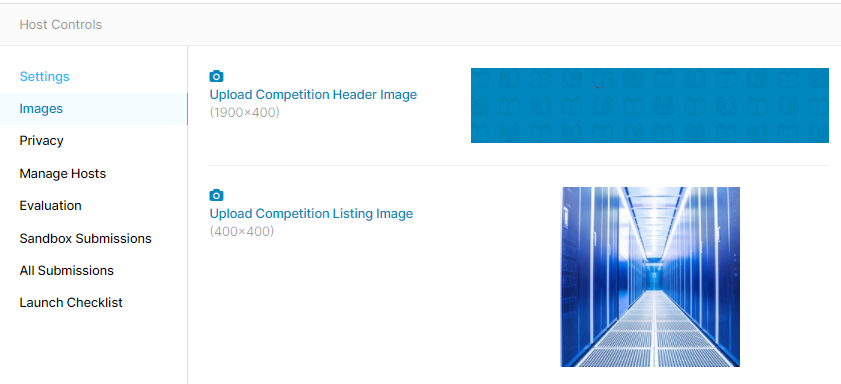


* A continuación, se describen las opciones de configuración que nos permite kaggle:
  + Host – Settings: donde se muestran las configuraciones básicas que podemos editar.

|  |  |
| --- | --- |
|  | -Titulo de la competición.  -URL de la competición.  -Una breve descripción.  -La organización del evento.  -La fecha límite en la que estará disponible la competición. |
|  | -La posibilidad de crear equipos.  -Máximos envíos diarios.  -Máximo número de envíos que aparecen en la clasificación |

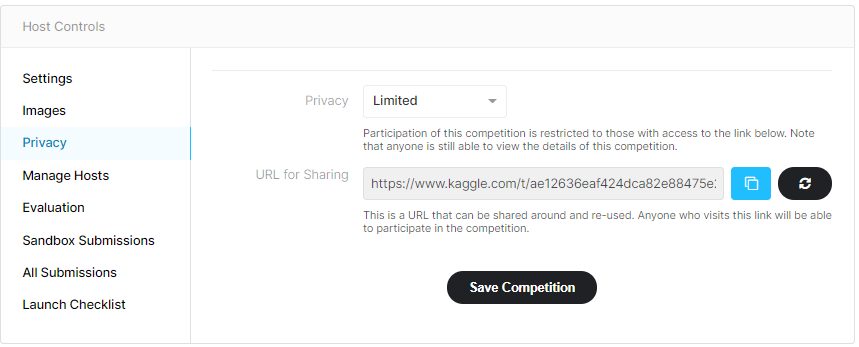
* Host – Images.

Nos permite cambiar las imágenes de muestra de nuestra competición

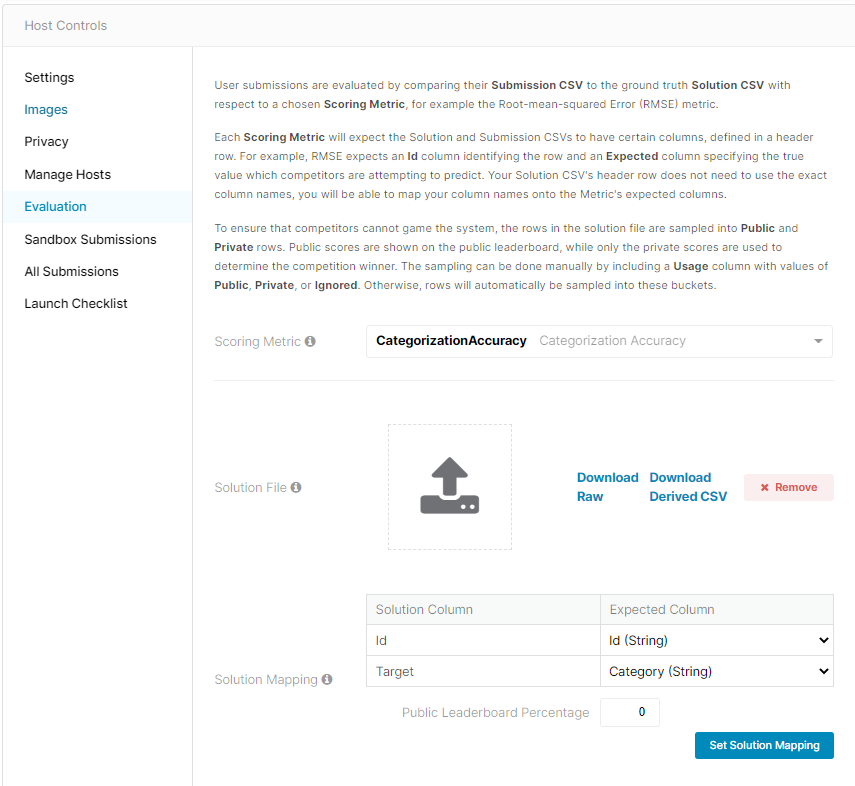


* Host – Privacy

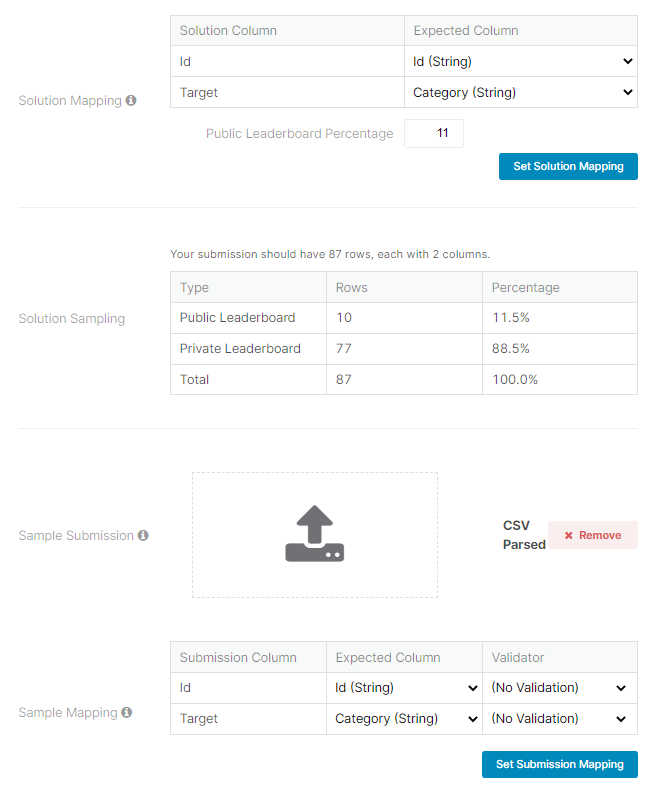
Podemos elegir entre hacer nuestra competición publica, para cualquier usuario, o limitada para restringir la participación solamente a aquellos que tengan la url.



* En el apartado de evaluación, vamos a especificar cómo tienen que ser los envíos que los alumnos tienen que realizar y que servirá para comparar sus predicciones contra el de la solución que hayamos preparado seleccionando una métrica de puntuación entre las disponibles como, por ejemplo, precisión en la categorización.



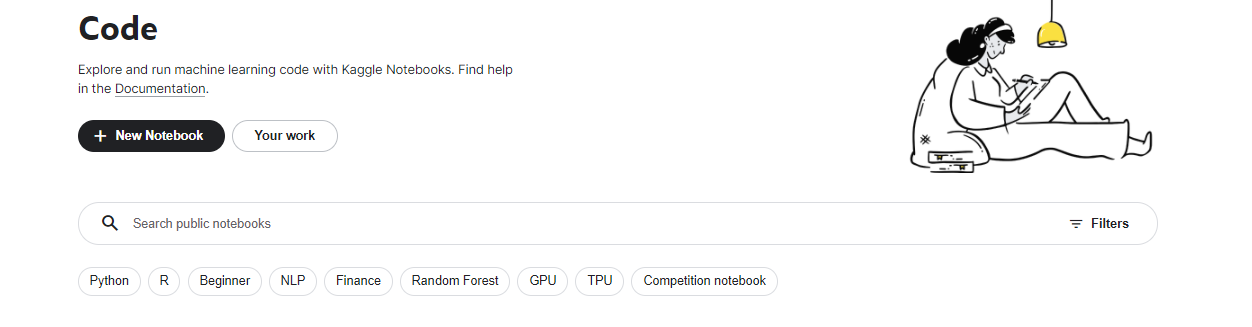
* Para poder evaluar cada CSV de solución y envío deberán de tener determinadas columnas, definidas en una cabecera que en la imagen puede verse en *Solution Mapping*, que ayuda a que las cabeceras de las columnas se le pueda asignar los nombres de las columnas esperadas de para el cálculo de la métrica.
* Una vez subida la solución indicamos el porcentaje del conjunto de test que se utiliza en la tabla de clasificación pública, en este caso un 11%, y dejamos el 89% del conjunto restante en la tabla de clasificación privada que ocultaremos para que los usuarios no puedan jugar con los envíos. Esta será la tabla que decidirá los ganadores reales de la competición.



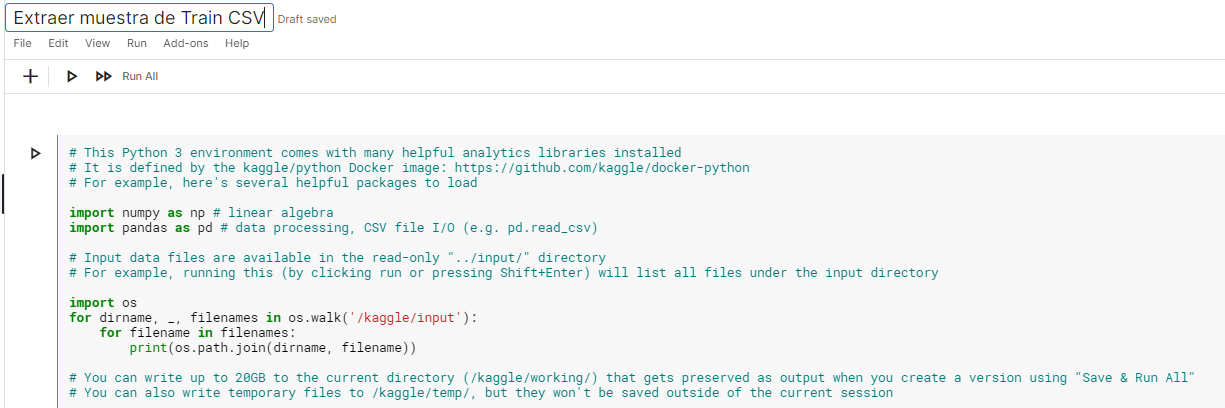
Como en este paso necesitamos especificar también las muestras de envío además del fichero csv de la solución, vamos a ver cómo podemos de manera rápida y sencilla extraer los mismos desde un dataset que ya exista en kaggle.

### 2.1. Cómo crear los conjuntos de datos necesarios para la competición desde los datasets ya existentes en Kaggle.

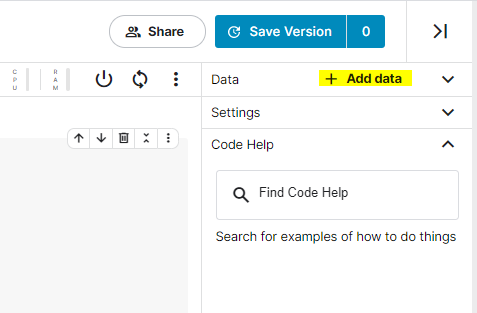
* En la ventana principal seleccionamos la opción code y +new Notebook



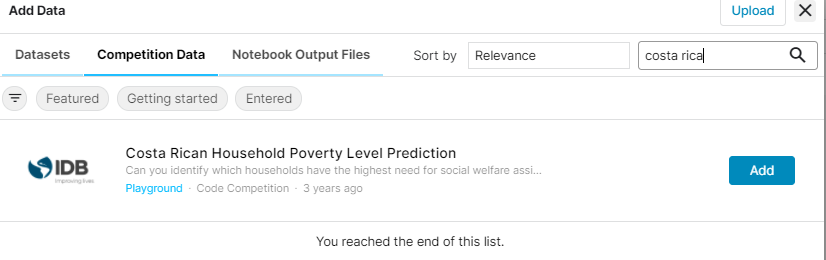
* Esto nos lleva a un cuaderno de Jupyter que nos aporta una interfaz muy sencilla, con una barra de menú, herramientas y con una celda inicial, que es la unidad básica de trabajo con Jupyter. En una celda, podremos escribir una o varias líneas de código y ejecutarla. También podemos escribir texto plano y eliminar la celda del flujo de ejecución del cuaderno o hacer que se muestre de manera diferente si queremos que funcione como título, separador de sección, etc.



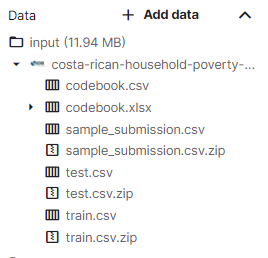
* En la esquina superior derecha vemos que nos permite traernos datos a nuestro notebook. Le damos a añadir datos:



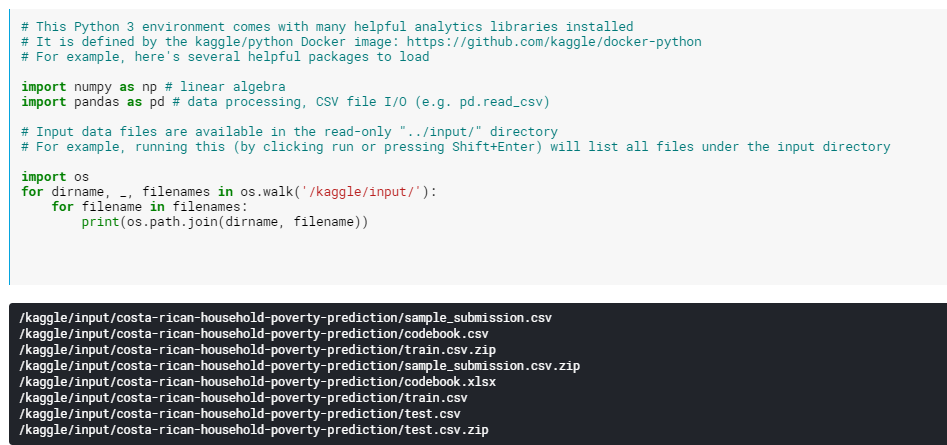
* Buscamos el dataset que queramos utilizar en la competición y le damos a añadir.



* Realizado este paso, veremos que contamos con los conjuntos de datos ya cargados en nuestro cuaderno.

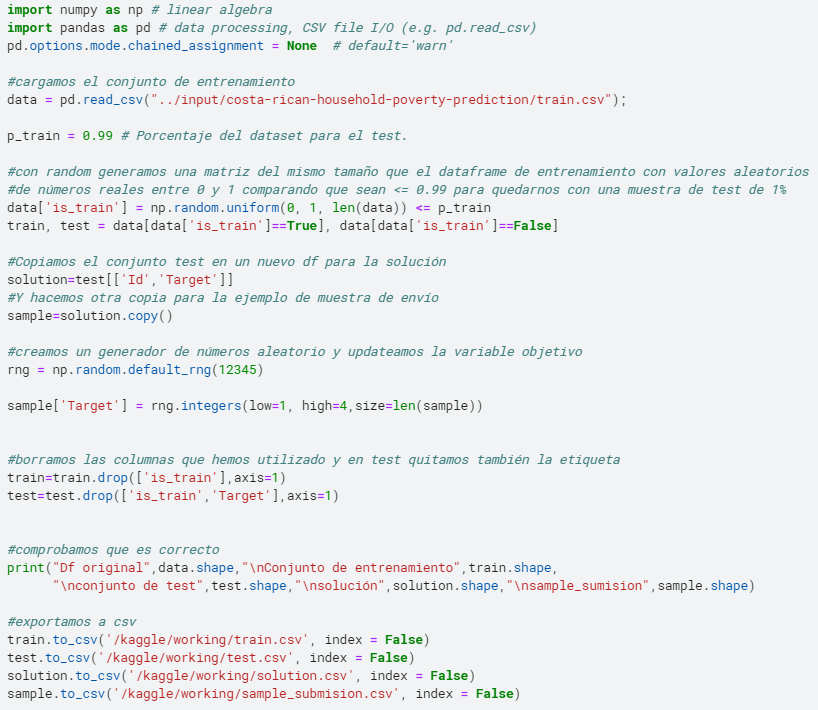


* Si ejecutamos el código de muestra, vemos que estos datos son accesibles desde nuestra celda de ejecución.

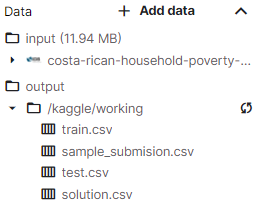


* Borramos el código de muestra y copiamos el que viene aquí adjunto:





* Tras la ejecución vemos que se han creado los conjuntos de datos que necesitamos para continuar con nuestra competición:

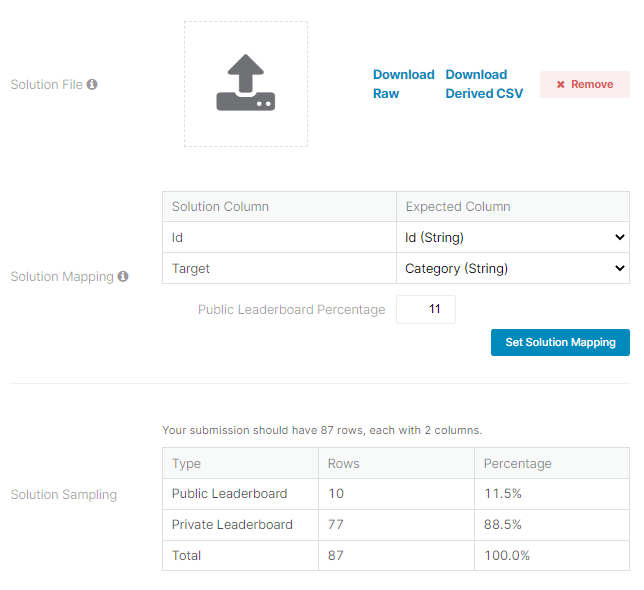


* Por último, nos descargamos el conjunto de datos para poder utilizarlo en la fase final de puesta a punto de nuestra competición.

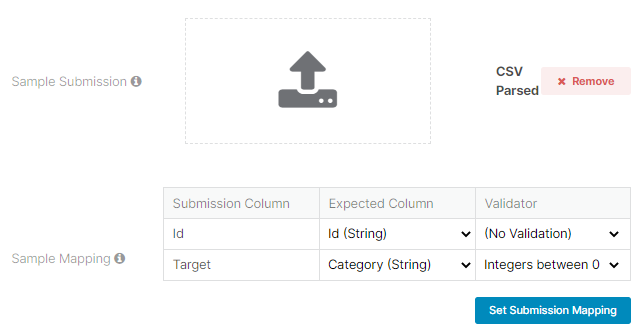
### 2.2. Lanzamiento de competición.

Con los conjuntos de datos ya creados, continuamos completando las opciones de evaluación.

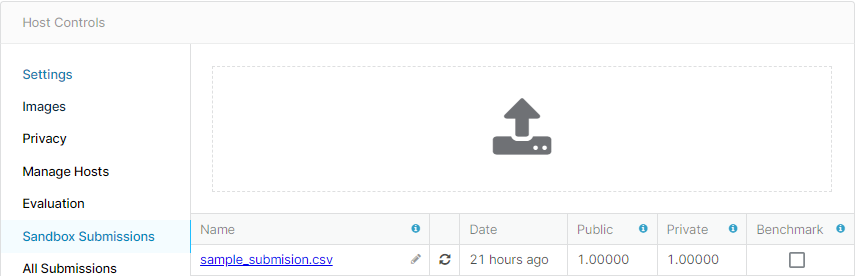
* Subimos el csv de solución e indicamos qué muestra de la misma servirá para la tabla de clasificación pública.



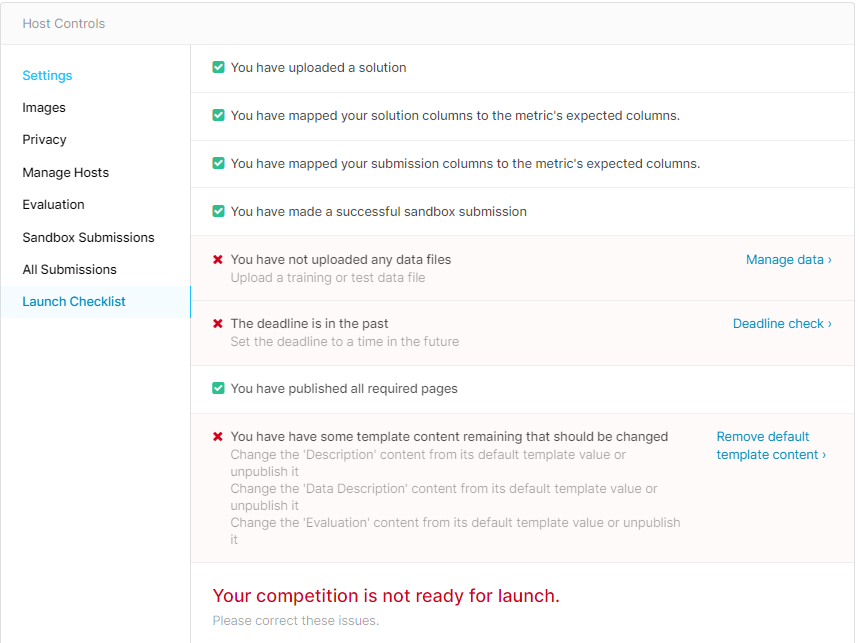
* Establecemos el mapping entre la solución y la columna esperada como ya se explicó anteriormente.
* A continuación, subimos la muestra de datos para que los usuarios sepan cómo tienen que enviar los datos indicando el mapeo y un validador de los mismos.



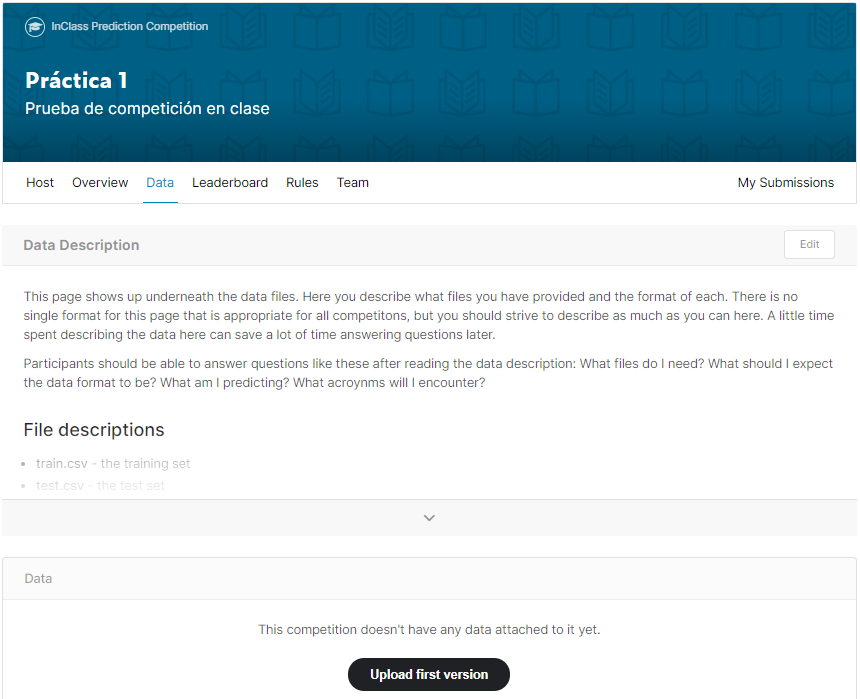
* Continuando con las opciones tenemos un apartado para probar los envíos



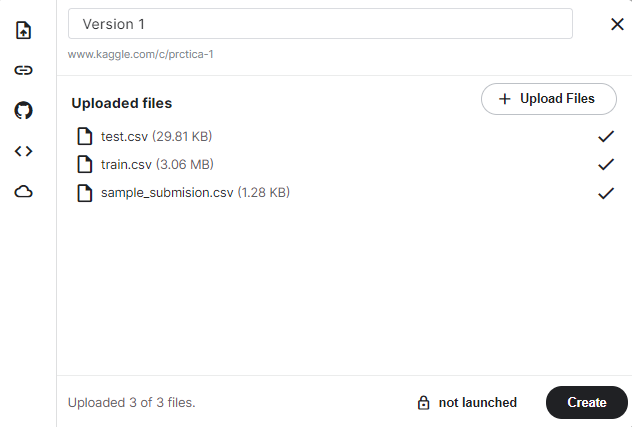
* Por último, si queremos lanzar en este estado la competición nos dará los siguientes errores:



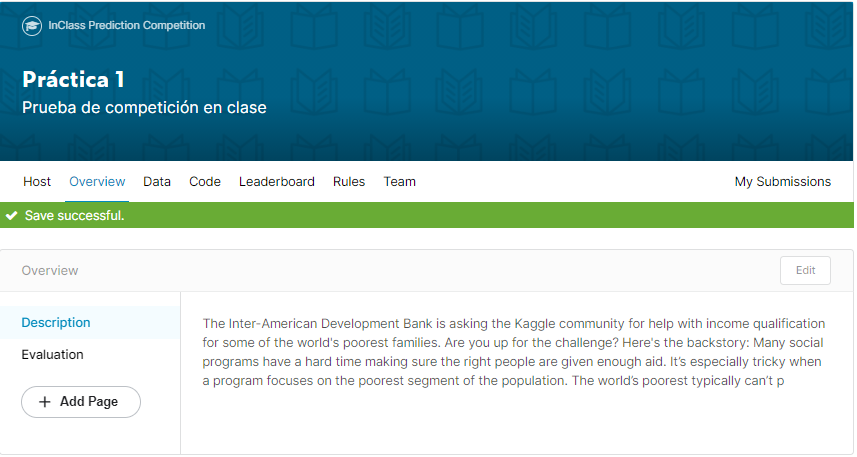
* Tenemos que subir primero los datos de entrenamiento y test, para ello vamos a DATA y subimos los conjuntos que hemos creado en el apartado 2.1. También nos pedirá añadir las descripciones de los datasets.



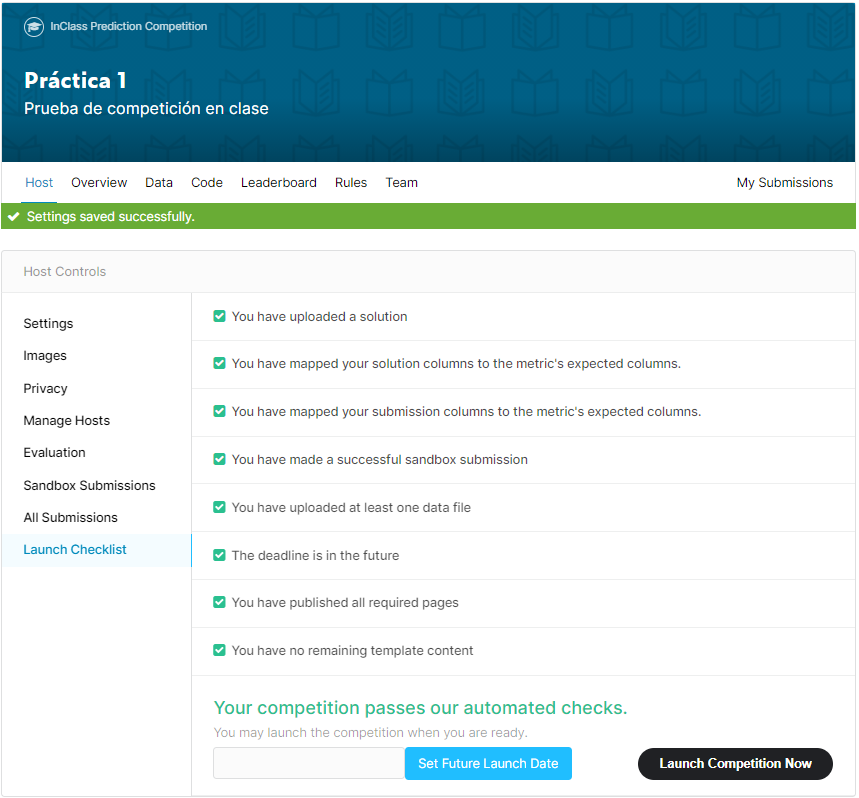
* Subimos los archivos:



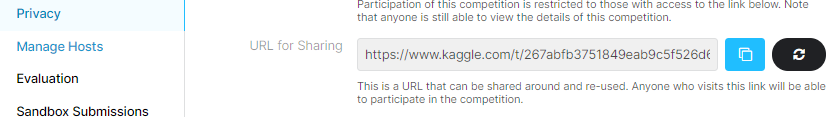
* Modificamos el Overview de la competición.



* Y por último añadimos un deadline, desde Host🡪 Settings para el límite de la competición.
* Volvemos a hacer la comprobación y ahora ya comprobamos que está todo ok.



* Una vez completados los pasos anteriores, se deberá enviar la url de la competición a: [clopez@teralco.com](mailto:clopez@teralco.com)

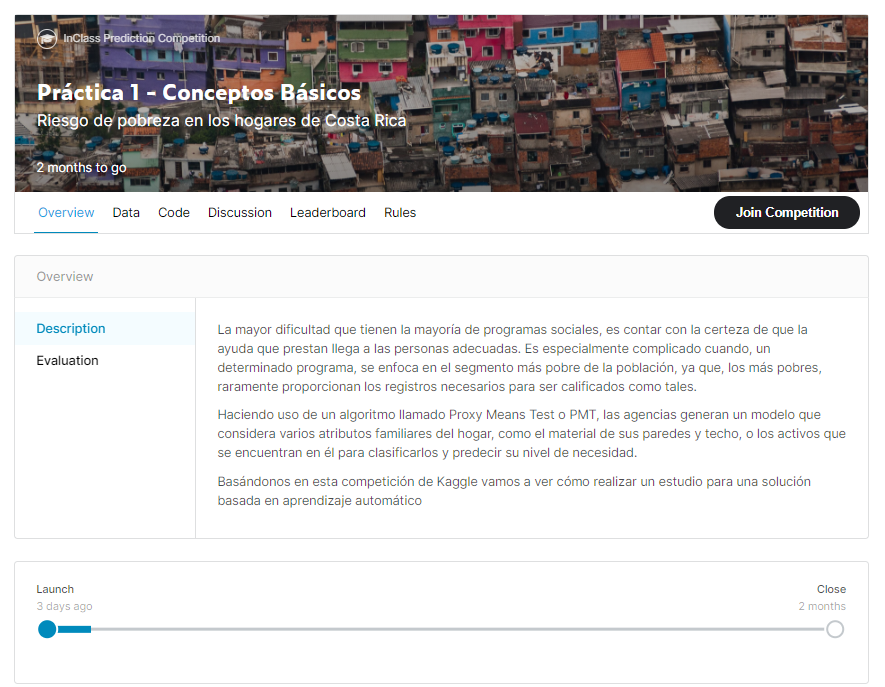


## 3. Cómo participar en una competición

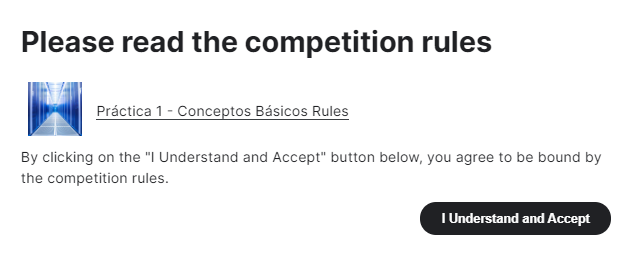
Para el objeto de esta práctica, se ha creado una competición donde el alumno deberá participar a través del siguiente enlace:

<https://www.kaggle.com/t/0f4cdf9e4ba545f0a698803c29d8e24e>

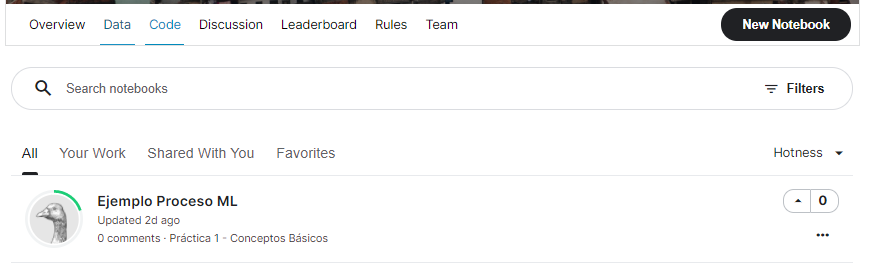
* Una vez se accede, le damos a *Join Competition* para unirnos a la competición.



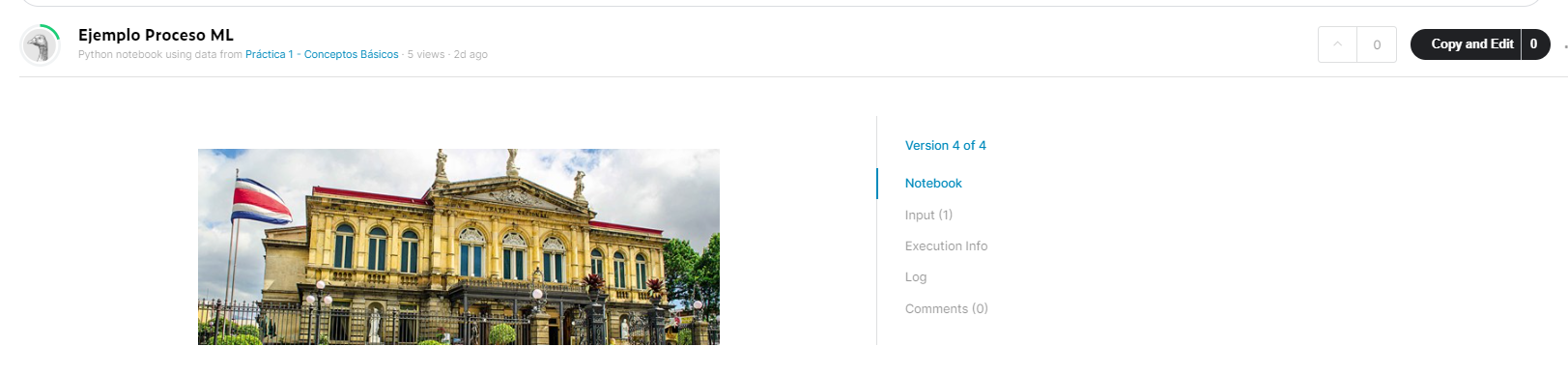
* Le damos a aceptar:



* Una vez hemos aceptado ya estaremos dentro de la competición. Para empezar, hay que seleccionar la opción Code y abrir el notebook que se muestra en la siguiente imagen:



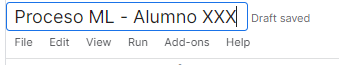
* Una vez se nos muestra el notebook de ejemplo le damos a copiar y editar en la esquina superior derecha.



* Ahora ya solo queda ir ejecutando las celdas para entender el proceso y completar las partes marcadas como ejercicio:



* Una vez completado los ejercicios de la parte de visualización, se ha de ejecutar el modelo del ejercicio para generar una predicción y enviarla para que sea valorada tal y como se detalla en el notebook.
* Por último, se cambia el nombre del notebook.



* Y se comparte de manera privada añadiendo al colaborador “carloscursobd” para que pueda ser evaluada, tal y como se muestra en la siguiente imagen.

