

### Actividad 3.7 - Predicción de Riesgo de derrumbamiento - Terremotos

El objeto de esta actividad es participar en la competición de ofrecida de la web de DrivenData denominada: Richter's Predictor: Modeling Earthquake Damage.

**Título:** Richter's Predictor: Modeling Earthquake Damage

**Url:** <https://www.drivendata.org/competitions/57/nepal-earthquake/data/>

La actividad consiste en subir a dicha web un fichero csv con la estimación del nivel de daño provocado en un edificio/vivienda tras un terremoto (1 - low damage, 2 - medium amount of damage, 3 - complete destruction) que hayamos obtenido al aplicar **al menos tres modelos de árboles y uno de SVM**.

Podrás observar que se pueden realizar hasta un máximo de 3 subidas diarias y la propia web realizará una valoración de tu solución. La valoración se realiza utilizando el **micro averaged F1 score** (*variante* de **F1 score** – Ver D18 del UT3 - Algoritmos y herramientas para el aprendizaje(I)\_v2.pdf) como criterio de valoración de calidad. Esta técnica es recomendable utilizarla cuando los datos no están balanceados.

**Título:** sklearn.metrics.f1\_score – (Fijarse en *average= 'micro'*)

**Url:** [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html)

Richter's Predictor: Modeling Earthquake Damage

HOSTED BY DRIVENDATA

10 MONTHS LEFT

Glory!

LEADERBOARD

DATA DOWNLOAD

SUBMISSIONS

TEAM

DISCUSSION

OFFICIAL RULES

Data

This competition is for fun, and the data is available for use outside of DrivenData. If you use the dataset for projects, we encourage you to share your work on the submissions page (click \*\*), the forum or on [Twitter](#)! If you publish work, please cite [the DrivenData platform page](#).

File	Description
<a href="#">Submission Format</a>	The submission format.
<a href="#">Test Values</a>	The test values.
<a href="#">Train Labels</a>	The training labels.
<a href="#">Train Values</a>	The training features.

Based on aspects of building location and construction, your goal is to predict the level of damage to buildings caused by the 2015 Gorkha earthquake in Nepal.



## Richter's Predictor: Modeling Earthquake Damage

HOSTED BY DRIVENDATA

regression is sometimes described as an problem somewhere in between classification and regression.)

To measure the performance of our algorithms, we'll use the **F1 score** which balances the **precision** and **recall** of a classifier. Traditionally, the F1 score is used to evaluate performance on a binary classifier, but since we have three possible labels we will use a variant called the **micro averaged F1 score**.

$$F_{micro} = \frac{2 \cdot P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}}$$

where

$$P_{micro} = \frac{\sum_{k=1}^3 TP_k}{\sum_{k=1}^3 (TP_k + FP_k)}, \quad R_{micro} = \frac{\sum_{k=1}^3 TP_k}{\sum_{k=1}^3 (TP_k + FN_k)}$$

and  $TP$  is True Positive,  $FP$  is False Positive,  $FN$  is False Negative, and  $k$  represents each class in 1, 2, 3.

In Python, you can easily calculate this loss using `sklearn.metrics.f1_score` with the keyword argument `average='micro'`. Here are some references that discuss the micro-averaged F1 score further:

Consideraciones a tener en cuenta (**Leer la rúbrica**):

- Dado que el dataset que nos ofrece el reto contiene un número de filas muy elevado de instancias/filas para realizar el ejercicio, se hace necesario el realizar una selección de un conjunto de ellas que consideres oportuno. Sea cual sea la técnica/herramienta o criterio que propongáis lo has de justificar. Este punto se valorará mejor cuanto menos aleatorio sea.
- Aplicar al menos tres modelos de árboles y uno de SVM entre los vistos en clase.
- Realizar pruebas de hiper-parametrización con las dos técnicas explicadas: GridSearch y RandomSearch.
- Es fundamental el utilizar la herramienta de los dendogramas para determinar qué conjunto de características nos conviene seleccionar para realizar la predicción.
- Utilizar la librería (**Lazy Predict**) indicada en la siguiente publicación con el objeto de determinar si realmente puede resultar interesante a la hora de decidir el modelo a utilizar en la resolución de un problema.

**Título:** lazypredict 0.2.12

**Url:** <https://pypi.org/project/lazypredict/>

**Título:** Do you need to build a ML classifier but don't know what model to use? Try "Lazy Predict" to get an idea of the most promising models

**Url:** [https://www.linkedin.com/posts/agostino-calamia\\_mlops-data-datascience-activity-7044956022095945730-mwVB?utm\\_source=share&utm\\_medium=member\\_desktop](https://www.linkedin.com/posts/agostino-calamia_mlops-data-datascience-activity-7044956022095945730-mwVB?utm_source=share&utm_medium=member_desktop)

## La rúbrica a utilizar para evaluar

Criterios para valorar Proyecto/Reto	
Actividad 3.7 - Predicción de Riesgo de derrumbamiento - Terremotos	
Peso %	Tareas (Se evalúa entre 0 y 10)
2	Utilizar el <u>drive/github</u> como origen de ficheros para la importación del <u>dataset</u> .
5	Importación del <u>dataset</u> : Preparación de los datos: Normaliza, ajusta la calidad de los datos
5	Selección de <u>características</u> : Explora herramientas gráficas o no gráficas, que <u>no sean los dendogramas</u> , para la elección de las <u>características</u>
5	Selección de <u>características</u> : Utiliza <u>dendogramas</u> para la elección de las características
5	Además de la división de los datos de <u>train</u> y test, incorpora la utilización de datos de validación.
12	Entrenamiento: Trabaja con tres modelos de árboles - Desarrolla las diversas pruebas propuestas para la selección y <u>justifica</u> el criterio de calidad para la selección del modelo que mejores resultados ofrece. Utiliza <u>Cross Validation</u>
13	Entrenamiento: Elegir un modelo de regresión o SVM - Desarrolla las diversas pruebas propuestas para la selección y <u>justifica</u> el criterio de calidad para la selección del modelo. Utiliza <u>Cross Validation</u> y pruebas de <u>hiperparámetros</u> ( <u>GridSearch</u> y <u>RandomSearch</u> )
10	Entrenamiento: Utilizar la librería <u>Lazy Predict</u> para determinar otros posibles modelos a utilizar para la resolución del problema, elije uno de los modelos propuestos y compara la predicción con el modelo de regresión o <u>SVM</u> elegido
10	Predicción: Utiliza herramientas gráficas para ayudar a entender la precisión de los resultados obtenidos
5	Predicción: Describe con claridad una valoración de los resultados obtenidos.
10	<u>Submit</u> del fichero con la predicción y captura de la valoración /posicionamiento obtenido en la competición
5	Propone soluciones creativas e innovadoras
3	Registra en el <u>pdf</u> final, un cuaderno de bitácora/seguimiento donde se muestra los <u>submits</u> realizados, explicando los ajustes o mejoras que han motivado cada subida o un grupo de <u>submits</u>
5	El <u>pdf</u> final tiene una portada., utiliza un índice, apartado de conclusiones y referencias (web). Se hace <u>mención</u> a referencias externas, no recogidas en el material suministrado.
5	Comenta con claridad cada uno de los pasos realizados.
Total	100

## Formato de entrega

- Entregar un fichero en un Archivo PDF con capturas del código y resultados obtenidos, así como **la url de GitHub y Google Colab donde has publicado el código.**
- Nombrar el archivo siguiendo el siguiente patrón:

SNS\_ACT3\_7\_NombreApellidos.pdf