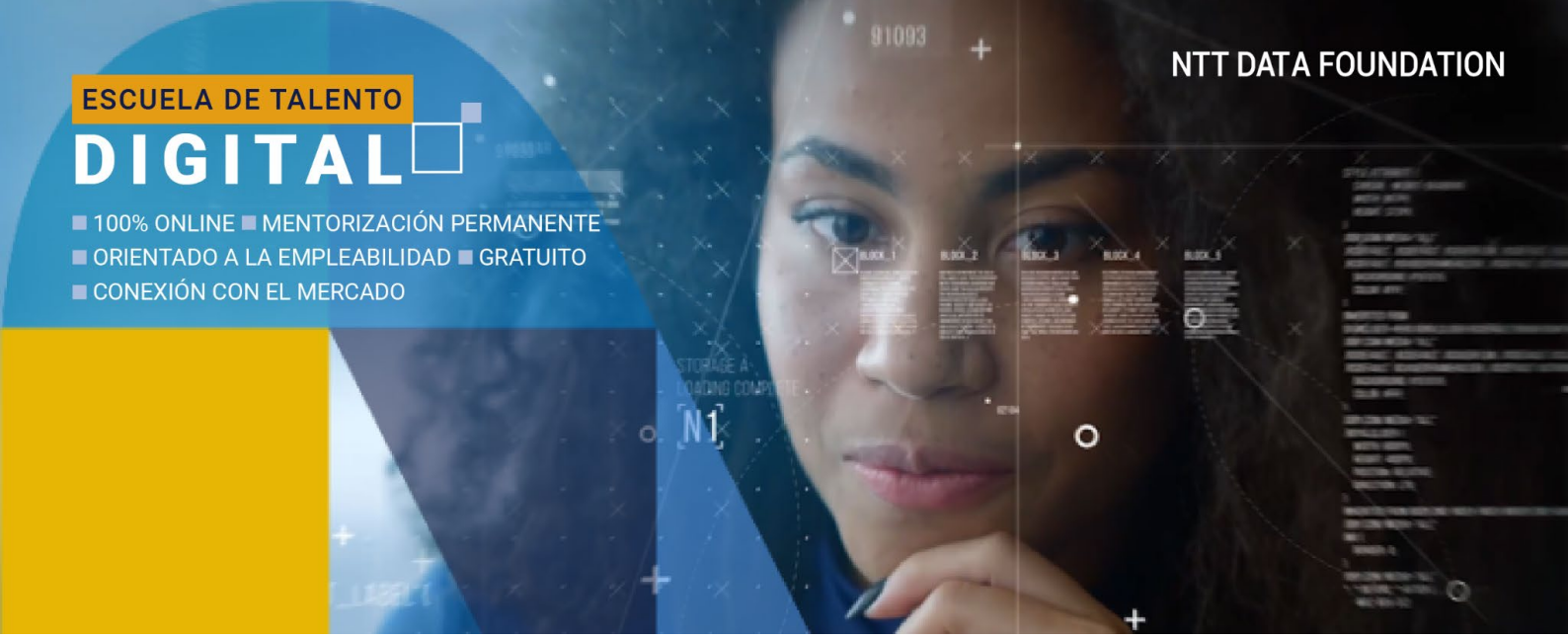


ESCUELA DE TALENTO DIGITAL

NTT DATA FOUNDATION

RETO GRUPAL ÁREA 3

ÍNDICE

1. INTRODUCCIÓN

Para superar este reto grupal, tendréis que ir resolviendo una serie de ejercicios que os vamos a proponer en este documento.

1.1. ¿Cómo entregáis vuestros ejercicios?

Tendréis que preparar un documento pdf y subirlo a la plataforma en el espacio habilitado para ello. No es necesario que todos los componentes del grupo subáis el documento, con que lo suba uno de vosotros es suficiente.

1.2. ¿Qué debe contener el documento pdf?

Este documento deberá contener el código propuesto para resolver cada uno de los ejercicios, pero, además, también debe contener una explicación de cómo habéis llegado a obtener esa solución, que debe ser conjunta y aprobada por todos los miembros del grupo.

Como durante el desarrollo de la actividad van a surgir diferentes propuestas, queremos que las documentéis, es decir, que cuando expliquéis cómo habéis llegado al resultado final, también tenéis que explicar qué otras alternativas había y quién las ha propuesto.

Por ejemplo, imaginad un grupo de 5 alumnos (alumno1, alumno2, alumno3, alumno4 y alumno5) resolviendo el ejercicio 1. La propuesta de resolución del ejercicio 1 debería ser algo como esto:

Después de leer el enunciado, entendimos que lo que se solicitaba era hacer ……

Durante el proceso, el estudiante1 propuso llegar a la solución de la siguiente manera…… pero al estudiante2 y al estudiante3 les pareció mejor hacer …….. y todos estuvimos de acuerdo.

Por todo esto, proponemos esta solución en la que estamos de acuerdo los 5 participantes:

CÓDIGO PROPUESTO

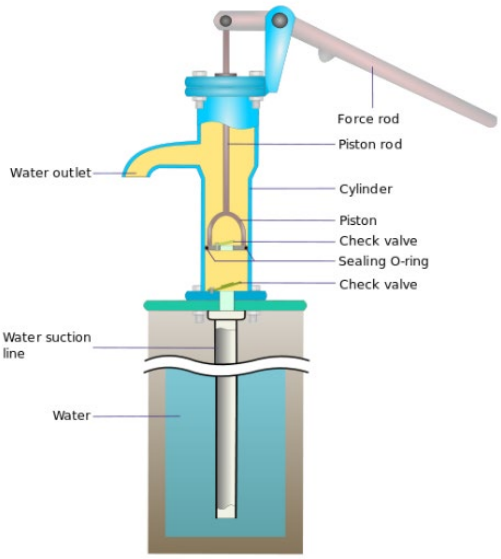
Con esto lo que queremos valorar es la participación de cada uno de vosotros durante el desarrollo del reto.

Cuidad también el formato en el que presentéis el documento, porque también se tendrá en cuenta.

Si tenéis cualquier duda, consultad al tutor a través de la plataforma.

2. CLASIFICACIÓN BINARIA DE PUNTOS DE AGUA, ENTRE FUNCIONALES Y NO FUNCIONALES

Este reto, que desarrollaréis a lo largo de todo el programa, en las diferentes áreas, tiene como objetivo que, al final del curso, logréis predecir qué bombas de Agua de Tanzania funcionan y cuáles no.



La información que se facilitará a lo largo de los distintos retos grupales tiene su base en un dataset obtenido en Driven Data, el cual presenta datos facilitados por el Ministerio de Agua de Tanzania, a cerca del estado de las distintas bombas de agua sobre las que tienen la competencia.

Una comprensión inteligente de qué puntos de agua fallarán puede mejorar las operaciones de mantenimiento y garantizar que las comunidades de Tanzania dispongan de agua limpia y potable.

Además, al finalizar el curso deberéis ser capaces de extraer toda la información posible de los datos facilitados y presentarla de la mejor manera posible, utilizando los gráficos y las herramientas de visualización vistas en clase.

Las variables que contiene este dataset y que, por tanto, servirán para los objetivos descritos anteriormente, son las siguientes:

- amount_tsh – carga estática total (cantidad de agua disponible, para el punto de agua).
- funder –quién financió el pozo.
- gps_height –altitud del pozo.
- installer –organización que lo instaló.
- longitude – coordenada GPS.
- latitude –coordenada GPS.
- wpt_name –nombre del punto de agua, si lo tiene.
- num_private –
- basin –cuenca hidrográfica.
- region –localización geográfica.
- population –población alrededor del pozo.
- public_meeting – True/False si es punto de reunión.
- recorded_by –grupo que introdujo este registro en los datos.
- scheme_management –quién opera el punto de agua.
- permit –si el punto de agua está permitido.
- construction_year –año de construcción.
- extraction_type –el tipo de extracción que utiliza el punto de agua.
- management_group – cómo se gestiona el pozo.

- payment_type – coste del agua.
- water_quality –calidad del agua.
- quality_group – calidad del agua.
- quantity_group – cantidad de agua que aporta el pozo.
- source_class – la fuente del agua.
- waterpoint_type_group – el tipo de punto de agua.

Partiendo de estas premisas, pasamos a enunciar los ejercicios de este tercer reto grupal que tendréis que resolver.

2.1. Ejercicio 1

Dado un fichero reto_agua.csv con los datos, realizad los siguientes puntos:

- **Cargad el csv**
- Mostrad los **primeros 5 datos**
- Realizad un **análisis exploratorio** de la estructura y los datos
- Extraed la información de la estructura del dataset para responder a las siguientes preguntas:
  - ¿Veis alguna **columna que no consideréis necesaria** para el modelo?
  - ¿Cuántos **datos totales** hay en dataset?
  - ¿Hay **valores nulos**? En ese caso, ¿qué columnas los tienen?
  - ¿Detectáis alguna columna que tenga **datos anómalos**? En ese caso, ¿cuáles?
- **Transformad** todas las variables objetos en categóricas o numéricas (se pondrán todas las filas nulas como una categoría más). Esto lo podéis hacer con un bucle, con apply, poniendo una a una las columnas, ...
- **Convertid** todas las columnas de columns_object en categóricas

NOTA: El código para obtener todas las columnas object es:

```
columns_object = df.loc[:, df.dtypes == object].columns
```

2.2. Ejercicio 2

Ahora, vamos a entrenar el modelo:

- Dividid los datos **en variable independiente** y **target**
- Dividid el modelo en un conjunto de datos para **el test (20%)** y otro para **el train (80%)** y **random_state=42**
- **Entrenad varios modelos** con los datos de train, **validadlo** con el test **y seleccionad el que mejor resultado obtiene**.

Una vez hecho esto, responded a las siguientes preguntas:

- ¿Qué **score** da el de entrenamiento y con el test?
- ¿Creéis que puede tener **sobreajuste** (overfitting) o **infraajuste** (underfitting)?

2.3. Ejercicio 3

Seleccionad las 21 variables que más influyen en la predicción y entrenad de nuevo el modelo. ¿Mejora?

Usadlas para sacar los scoring ['accuracy', 'precision', 'recall'] del conjunto de train:

- ¿Interpreta accuracy?
- ¿Interpreta precision?
- ¿Interpreta recall?
- ¿Predice mejor los positivos o los negativos?

2.4. Ejercicio 4

Validad la correlación con uno o más gráficos con las columnas ['amount_tsh', 'funder', 'gps_height', 'installer', 'longitude', 'latitude', 'num_private', 'basin','status_group'].

Después, **haced un gráfico**, el que consideréis adecuado, **para detectar outliers** en population y gps_height ¿alguno tiene outliers? De ser así, **eliminadlos** con el método de Inter cuartil con la columna o columnas con datos atípicos. ¿El modelo ha mejorado? Recordad que hay que volver a sacar los valores x e y (test y train).

Para terminar, **usad la búsqueda de hiperparámetro para ajustar al modelo** seleccionado (buscad en https://scikit-learn.org/ o en la página del modelo usado).