**DESCUBRIMIENTO DE LOS DATOS**

Se pretende analizar si la data disponible puede responder a las siguientes preguntas propuestas en [1] para poder tener un mejor entendimiento de la data a utilizar:

1. **¿Que data está disponible?**

La data disponible que se tiene es aquella relacionada a plantas y sus características/propiedades. Es de tipo biológica y tiene como propósito informar sobre las diferentes plantas y sus características

1. **¿Cuánta data está disponible?**

La data que se tiene disponible es de dos tipos: imágenes y archivos CSV. Por un lado, el conjunto de imágenes disponible para el entrenamiento y prueba del modelo consta de un total de cerca de 60K imágenes de plantas, por otro lado, se cuenta con un total de 3 archivos CSV que complementan la información acerca de las características de las plantas a analizar.

1. **¿Se tiene acceso a la verdad absoluta, los valores que estamos tratando de predecir?**

Si, se tiene un archivo CSV en donde se tiene una lista de las características/propiedades que se pretenden predecir de las imágenes. También se tienen valores deseados de estas características dentro del CSV a manera de ejemplo.

1. **¿En qué formato se encuentra la data?**

La data en imágenes se encuentra en formato JPEG y la data estructurada se encuentra en formato CSV

1. **¿En dónde residen los datos?**

Los datos son de propiedad de la página de [Kaggle](https://www.kaggle.com/competitions/planttraits2024/data?select=train.csv) y se encuentran disponibles para su descarga en la misma página. También se encuentran, por el momento, disponibles de manera local en la computadora.

1. **¿Cómo podemos acceder a los datos?**

Los datos son accesibles desde la página de [Kaggle](https://www.kaggle.com/competitions/planttraits2024/data?select=train.csv) para su descarga

1. **¿Qué campos/variables son los más importantes?**

Las variables “Y” a predecir son aquellas que se encuentran dentro de los archivos train.csv y test.csv, son aquellas que tienen de nombre “X[\*]\_mean”; también una variable importante dentro de los archivos de prueba y entrenamiento CSV es “id”, ya que esa es la llave mediante la cual podemos hacer el match de las imágenes de cada planta con sus propiedades dentro el archivo CSV.

Y, por supuesto, las imágenes de las plantas son los datos más importantes de este modelo.

1. **¿Cómo podemos unir las fuentes de diferentes datos?**

La variable clave o “llave” para unir las imágenes de cada planta con sus características/propiedades dentro del archivo CSV es la variable “id” dentro del CSV de entrenamiento y prueba, esta variable se une con el prefijo de cada imagen de entrenamiento y prueba, respectivamente.

1. **¿Qué métricas importantes son reportadas usando la data disponible?**

Las métricas de interés para este proyecto son aquellas que tienen la nomenclatura “X[\*]\_mean”, estas son las que al final tenemos que predecir para el proyecto y las que al final nos dirán si es que lo predicho con el modelo es lo correcto o no

1. **¿Cómo la data se acopla al método actual para completar la tarea asignada?**

El método actual con el que se plantea llegar a la solución del proyecto es mediante una red neuronal y bases de datos vectoriales. Ambos métodos están diseñados para trabajar con datos no estructurados y en este caso el tipo de datos disponibles con los que contamos son imágenes, por lo que son un tipo de dato no estructurado. Por otro lado, dentro de las bases de datos vectoriales, se pretende utilizar algoritmos de indexeo que nos permitan realizar búsquedas más eficientes y encontrar las similitudes entre imágenes, el cual es el resultado al que queremos llegar.

Por estas razones, es que la data disponible para la elaboración del proyecto se acopla con el método propuesto de solución

1. **Para cada uno de mis ejemplos/datos necesito tener el valor correcto, es decir, la categoría correcta que el modelo debe ser capaz de predecir (Ground Truth)**

Si, para cada dato/entrada se tiene el valor de “X[\*]\_mean” el cual se cuenta como la verdad absoluta de cada variable y la variable a predecir.

1. **Escribe todos los diferentes tipos de datos que tendrás disponible y evalúa si el modelo que quieres utilizar sea el intuitivamente correcto para predecir lo que quieres predecir. (Data Relevance)**

Para la elaboración del proyecto se cuenta con dos tipos de datos:

1. Imagen: Se cuenta con un total aproximado de 60K de imágenes de plantas sobre las cuales se plantea hacer el análisis correspondiente y sobre los cuales se plantea obtener las características necesarias para predecir valores futuros
2. Archivo CSV: Se cuenta con un archivo CSV con algunas otras características que nos van a ayudar a describir mejor las imágenes de las plantas y generar mejores resultados en el modelo.

Los métodos propuestos para la solución del proyecto se acoplan a los tipos de datos con los que se cuentan ya que están diseñados para tratar con data estructurada y no estructurada.

1. **¿Hay evidencia probada de que existe una conexión entre tu data disponible y lo que quieres lograr? (Data Relevance)**

Si, existen diferentes artículos de trabajo en los cuales se trabaja con imágenes y métodos de búsqueda y predicción de datos para generar resultados eficientes acerca de la clasificación o reconocimiento de imágenes, por lo que nos podemos respaldar en todos esos artículos.

**REFERENCIAS**

[1] <https://towardsdatascience.com/data-understanding-for-machine-learning-assessment-exploration-aca1aadc1cb6>