Universidad del Valle de Guatemala
Facultad de Ingeniería
Departamento de Ciencias de la Computación
CC3084 - Data Science - Ing. Lynette Garcia Perez

Análisis Exploratorio Proyecto 2

CGIAR Eyes on the Ground Challenge

Cristian Fernando Laynez 201281,
Juan Angel Carrera, 20593,
Sara Maria Paguaga 20634,
Guillermo Santos Barrios 191517

Guatemala, 7 de septiembre de 2023

Situación Problemática

En ciertas regiones de África, en donde predomina el clima seco y la falta de lluvia, los agricultores corren el riesgo de grandes pérdidas en sus cultivos debido a las sequías recurrentes. Con el propósito de reducir estos riesgos, ACRE África ayuda a los agricultores a gestionar el riesgo agrícola y lleva a cabo evaluaciones de pérdidas en base a imágenes de cultivos asegurados que son enviadas por los agricultores para evaluar la reclamación de seguros y determinar indemnizaciones adecuadas. No obstante, evaluar miles de imágenes de agricultores asegurados requiere mucho tiempo y ralentiza el proceso por lo que es necesario automatizar las liquidaciones de reclamos para hacer el proceso más eficiente y ágil.

Problema Científico

¿Cómo mejorar las soluciones existentes para predecir daños de la sequía y aumentar la precisión con la que un modelo predice daños a lo largo de varias estaciones?

Objetivos

Objetivo general

- Generar un modelo preciso capaz de predecir y analizar daños en los cultivos causados por la sequía en base a imágenes de cultivos.

Objetivos específicos

- Implementar algoritmos de aprendizaje de máquina para el análisis y clasificación de imágenes de cultivos.
- Evaluar la precisión de los modelos generados y seleccionar al modelo más eficiente.
- Comparar la precisión del modelo más eficiente generado con soluciones existentes para la clasificación de daños por sequía.

Descripción de los Datos

Operaciones de limpieza

Se evaluó si en los datos descargados había presencia de datos inválidos o duplicados. La revisión confirmó que la data está limpia, sin duplicaciones ni datos erróneos, reafirmando la calidad de la información para el análisis.

Vamos a verificar si existen filas repetidas con los mismos valores o no

In [132…
```
dataframe_train.duplicated().value_counts()
```

Out[132…
```
False    26068
dtype: int64
```

In [133…
```
dataframe_test.duplicated().value_counts()
```

Out[133…
```
False    8663
dtype: int64
```

In [134…
```
dataframe_sample.duplicated().value_counts()
```

Out[134…
```
False    8663
dtype: int64
```

Se puede ver que no existen filas repetidas

Vamos a verificar si hay data duplicada en cada una de las filas

In [135…
```
dataframe_train['filename'].value_counts().sort_values(ascending=False)
```

Out[135…
```
L427F01330C01S03961Rp02052.jpg    1
L415F01160C39S14146Rp41363.jpg    1
L341F00167C01S00324Rp14178.jpg    1
L1084F02394C39S13931Ip.jpg        1
L361F02347C01S10018Rp27925.jpg    1
                                 ..
L406F02369C01S06908Rp32150.jpg    1
L371F04405C20S11053Rp27732.jpg    1
L1150F01299C39S12449Rp33134.jpg   1
L342F01261C01S00366Rp01836.jpg    1
L406F00362C01S00614Rp06760.jpg    1
Name: filename, Length: 26068, dtype: int64
```

In [137…
```
dataframe_train['ID'].value_counts().sort_values(ascending=False)
```

Out[137…
```
ID_1S8OOWQYCB    1
ID_YBE047NM5J    1
ID_DBO3ZGI1GM    1
ID_ORZLWTEUUS    1
ID_84VIVATWWN    1
                ..
ID_MSIN1S49F3    1
ID_4OGOEC4FLT    1
ID_8YS6Z5J14V    1
ID_5FAHR85EJN    1
ID_ZVBF61EA0I    1
Name: ID, Length: 26068, dtype: int64
```

Análisis Exploratorio

A, B) Descripción de Variables y Observaciones Disponibles

	ID	filename	growth_stage	damage	extent	season
0	ID_1S8OOWQYCB	L427F01330C01S03961Rp02052.jpg	S	WD	0	SR2020
1	ID_0MD959MIZ0	L1083F00930C39S12674Ip.jpg	V	G	0	SR2021
2	ID_JRJCI4Q11V	24_initial_1_1463_1463.JPG	V	G	0	LR2020
3	ID_DBO3ZGI1GM	L341F00167C01S00324Rp14178.jpg	M	DR	60	SR2020
4	ID_ORZLWTEUUS	L1084F02394C39S13931Ip.jpg	V	G	0	SR2021
...	...	...	...	...	...	...
26063	ID_3II1SXC0ZO	L1084F03259C39S12149Rp41671.jpg	M	DR	30	SR2021
26064	ID_OE7OU9ZF4U	L406F04369C01S07190Rp22847.jpg	V	G	0	LR2021
26065	ID_20M531UIZZ	L134F00766C01S09784Rp26034.jpg	M	G	0	LR2021
26066	ID_BZBV2FH0KL	L1153F02464C01S00194Rp01561.jpg	F	G	0	SR2020
26067	ID_ZVBF61EA0I	L406F00362C01S00614Rp06760.jpg	V	G	0	SR2020

Número de observaciones (filas): 26,068
Número de variables (columnas): 6

Descripción de cada una de las variables
ID: Identificador único para cada observación (tipo objeto).
filename: Nombre del archivo de la imagen correspondiente a la observación (tipo objeto).

growth_stage: Etapa de crecimiento del cultivo (variable categórica, tipo objeto). Los posibles valores son:

- F: Flowering (Floración)
- M: Maturity (Madurez)
- S: Sowing (Siembra)
- V: Vegetative (Vegetativo)

damage: Tipo de daño observado (variable categórica, tipo objeto). Los posibles valores son:

- DR: Drought (Sequía)
- DS: Disease (Enfermedad)
- FD: Flood (Inundación)
- G: Good (Bueno)
- ND: Nutrient Deficient (Deficiencia de nutrientes)
- PS: Pest (Plaga)
- WD: Weed (Maleza)
- WN: Wind (Viento)

extent: Extensión del daño observado expresado en porcentajes con incrementos del 10% (variable numérica, tipo int64).

season: Temporada de la observación (variable categórica, tipo objeto). Los posibles valores son:

- LR2020
- LR2021
- SR2020
- SR2021

C) Resumen de las variables

Resumen de la variable numérica 'extent'

- Conteo: 26,068
- Media: ~7.10%
- Desviación estándar: ~18.61%
- Mínimo: 0%
- 25% (Primer cuartil): 0%
- 50% (Mediana): 0%
- 75% (Tercer cuartil): 0%
- Máximo: 100%

Tablas de frecuencia para las variables categóricas

Variable 'growth_stage'

- V (Vegetative): 10,015 observaciones
- M (Maturity): 6,664 observaciones
- F (Flowering): 6,164 observaciones
- S (Sowing): 3,225 observaciones

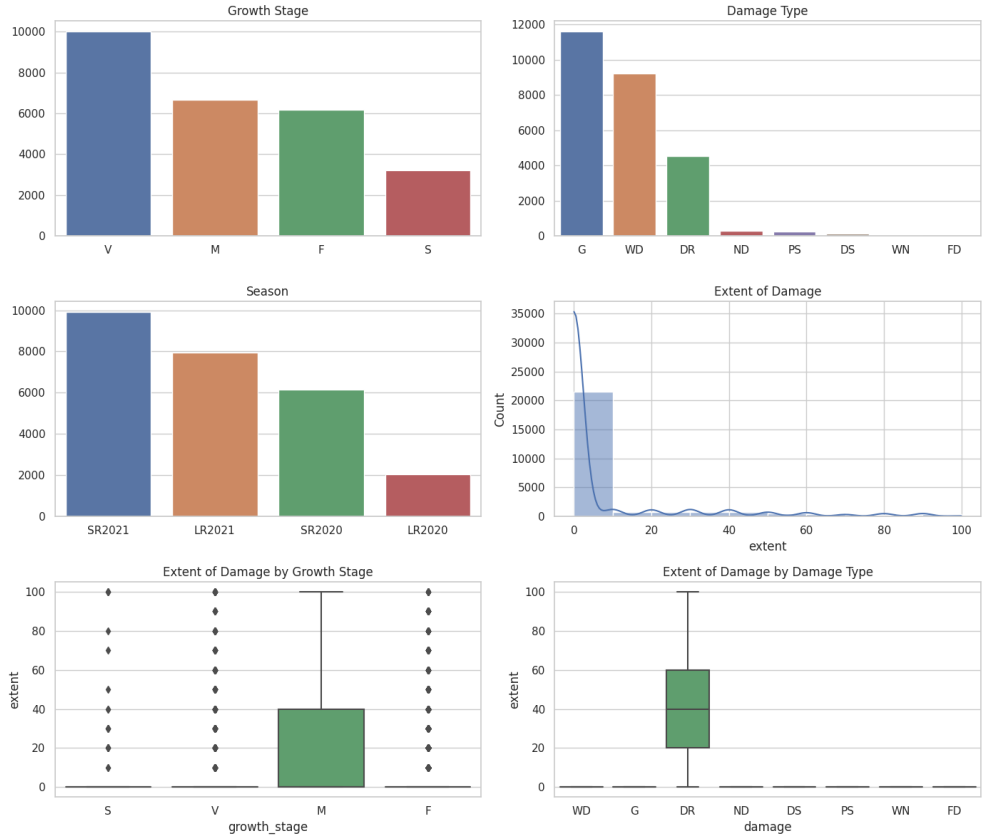
Variable 'damage'

- G (Good): 11,623 observaciones
- WD (Weed): 9,238 observaciones
- DR (Drought): 4,516 observaciones
- ND (Nutrient Deficient): 272 observaciones
- PS (Pest): 254 observaciones
- DS (Disease): 115 observaciones
- WN (Wind): 37 observaciones
- FD (Flood): 13 observaciones

Variable 'season'

- SR2021: 9,927 observaciones
- LR2021: 7,945 observaciones
- SR2020: 6,163 observaciones
- LR2020: 2,033 observaciones

D) Gráficos



Los gráficos anteriores nos proporcionan una visión detallada de cada variable y cómo se relacionan entre sí.

Análisis Individual de Variables

Growth Stage (Etapa de crecimiento)

La mayoría de las observaciones están en la etapa vegetativa (V), seguida por la etapa de madurez (M) y la etapa de floración (F).

Damage Type (Tipo de daño)

La categoría "Good" (sin daño) tiene la mayoría de las observaciones, seguida por "Weed" (presencia de maleza) y "Drought" (sequía).

Season (Temporada)

La temporada con más observaciones es SR2021, seguida de cerca por LR2021.

Extent of Damage (Extensión del daño)

La mayoría de las observaciones tienen un porcentaje de daño del 0%, lo que indica que hay muchas observaciones sin daño. El histograma muestra una distribución muy sesgada hacia la derecha.

Análisis Cruzado de Variables

Extent of Damage by Growth Stage (Extensión del daño por etapa de crecimiento)

Se observa que la mediana del daño en todas las etapas de crecimiento es 0%. Sin embargo, hay una presencia significativa de outliers, especialmente en las etapas de madurez (M) y vegetativa (V).

Extent of Damage by Damage Type (Extensión del daño por tipo de daño)

Aquí podemos ver cómo varía el daño en función del tipo de daño. La categoría "Good" tiene una mediana de 0%, como era de esperar. La categoría "Drought" (DR) muestra una mayor variabilidad en la extensión del daño, con varios casos extremos.

Hallazgos y conclusiones

- La falta de observaciones en la temporada LR2020 puede causar que el modelo no sea del todo preciso al predecir con esos parámetros.
- La mayoría de las observaciones con daño son a causa del crecimiento de maleza.
- Hay una gran cantidad de datos sin daño, lo que podría afectar al análisis y al entrenamiento de modelos predictivos en el futuro.
- Utilizar un proceso de aumentación de datos será útil para mejorar la capacidad del modelo para predecir.
- Sería útil investigar más a fondo las relaciones entre las variables categóricas y la extensión del daño, posiblemente a través de análisis de correlación o pruebas estadísticas para determinar las relaciones significativas.

Enlace de google docs:
https://docs.google.com/document/d/1F_riIwepfNdkYA-ZLMZLXnVmA2gy45nowlHumznT2CM/edit

Enlace del repositorio:
https://github.com/MGonza20/Proyecto2_DataScience