EII 7446 Project Entrega Final

Daniela Jofré Quiroz, Richard Lobos Araya, Mario Ríos Serrano, Sergio Sánchez Inostroza

En el contexto de desarrollar una herramienta que permitiera tanto a procuradores como a personas no relacionadas al arte, reconocer una obra de arte, su tipo, su autor, año y material, a través de las imágenes se abordó el Rijksmuseum Challenge 2014. El objetivo es encontrar la categoría a la que pertenece cada imagen según los atributos que se obtienen del procesamiento de imágenes, a través de alguna aplicación de Data Science.

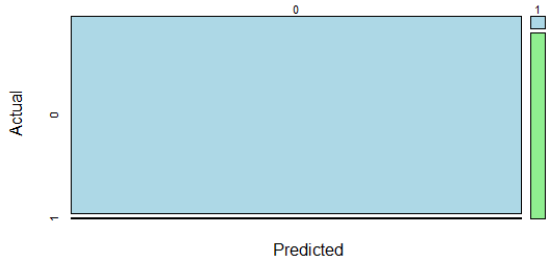
A continuación, se presenta un modelo de machine learning, “Decision Tree Classifier”, que dado las dimensiones de una obra de arte, el material y la técnica, indica si esa obra corresponde a una pintura.

Recordando que la base de datos quedó de 112.000 obras de arte con 19 atributos (variables), de los cuales para la clasificación se utilizan: largo, ancho, profundidad, material y técnica. Las dimensiones se consideran en centímetros; respecto a material y técnica se castean como factores, que tienen 241 tipos de materiales y 385 tipos de técnicas.

El conjunto de entrenamiento corresponde al 60% de los datos (67.224 obras de arte), y el de testeo corresponde al 40% (44815 obras de arte). También el outcome conocido como pintura, también es casteado como factor con dos categorías 1 es pintura, 0 no es pintura.

Luego del proceso de entrenamiento, testeo se realiza la clasificación predictiva con el modelo. Para comprobar el desempeño del modelo se realiza una matriz de confusión obteniendo los siguientes resultados.

Matriz de Confusión




```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 43301   181
         1    81  1252

               Accuracy : 0.9942
                 95% CI : (0.9934, 0.9948)
    No Information Rate : 0.968
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9023

 Mcnemar's Test P-Value : 9.581e-10

            Sensitivity : 0.9981
            Specificity : 0.8737
         Pos Pred Value : 0.9958
         Neg Pred Value : 0.9392
             Prevalence : 0.9680
         Detection Rate : 0.9662
   Detection Prevalence : 0.9703
      Balanced Accuracy : 0.9359

       'Positive' Class : 0
```

El modelo en comparación con el grupo de testeo, clasifica de forma correcta 43.301 obras de arte como no pinturas y 1.252 obras de arte como pinturas. Mientras que hay un error de 81 obras de arte que son clasificadas como pinturas y no lo son y 181 pinturas que erróneamente son clasificadas como otras obras de arte. Recordar que la categoría de otras obras de arte comprende, cerámicas, esculturas, fotografías, dibujos, entre otros. El modelo clasifica correctamente el 99.42% de las instancias.

Respecto de los otros Challenges, quedaron pendientes, debido a dificultades con la estructuración de la base de datos obtenida vía código MatLab de los 120.000 archivos XML. Se desarrollaron tres versiones de BD para poder implementar el modelo de clasificación, BD_Arte, BD_Arte_versión2 y BD_ArteOF2. La última es la versión que se utilizó para el modelo, dado que en ella se logró desagregar los atributos de dimensiones, tipo de material y técnica. Las tres bases y líneas de código asociadas se encuentran en el github: https://github.com/Maerios11xx/EII7446.