CLASIFICADOR SPAM/HAM CON ÁRBOL DE DECISIÓN (CART)

CESAR AGUIRRE HURTADO

LINK GITHUB: https://github.com/cesard2003/Arbol-Cart

UNIVERSIDAD DE CUNDINAMARCA
ALEXANDER ESPINOSA
MACHINE LEARNING
SEPTIEMBRE

Introducción

En este informe presento el desarrollo de un clasificador de correos electrónicos para diferenciar entre SPAM y HAM usando un Árbol de Decisión, específicamente el algoritmo CART implementado con scikit-learn. La idea principal es automatizar la detección de correos no deseados y analizar qué características del correo son más importantes para la clasificación.

Para evaluar correctamente el desempeño del modelo, se calculan métricas como Accuracy, F1 Score y Z-score, y se repite la ejecución del modelo 50 veces con diferentes divisiones de entrenamiento y prueba. Esto permite ver cómo cambia el rendimiento del modelo según la muestra utilizada y asegura que los resultados sean confiables y estables. Además, se analiza la importancia promedio de las características, lo cual ayuda a entender qué información del correo influye más en la clasificación.

Procedimiento

Primero, cargamos el dataset dataset_correos_1000_instancias.csv y separamos los datos en variables predictoras (X) y la variable objetivo (y). La variable objetivo indica si el correo es “Ham” o “Spam”, y se codifica como 0 y 1 respectivamente, porque los modelos de scikit-learn requieren valores numéricos.

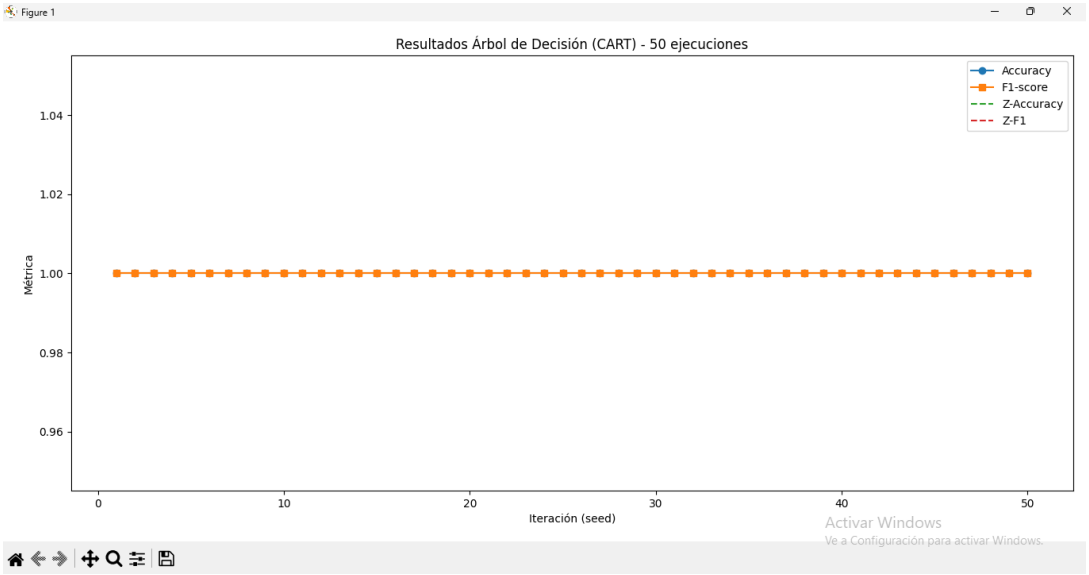
Dado que todas las columnas del dataset son categóricas, se aplicó One-Hot Encoding para convertir cada categoría en columnas binarias, permitiendo al árbol de decisión procesarlas correctamente. Para mantener un flujo organizado y reproducible, usamos un pipeline que integra el preprocesamiento y el modelo en un solo paso.

El modelo elegido es un Árbol de Decisión CART, que divide los datos en nodos basados en la característica que mejor separa las clases. Se ejecutó el modelo 50 veces con diferentes semillas para obtener distintas particiones de entrenamiento y prueba. Esto permite evaluar la estabilidad del modelo y entender cómo pequeñas variaciones en los datos afectan los resultados.

Para cada ejecución se calculan Accuracy, F1 Score y Z-score, y se almacenan las matrices de confusión y la importancia de las características. Estas métricas permiten analizar el desempeño del modelo y la relevancia de cada atributo.

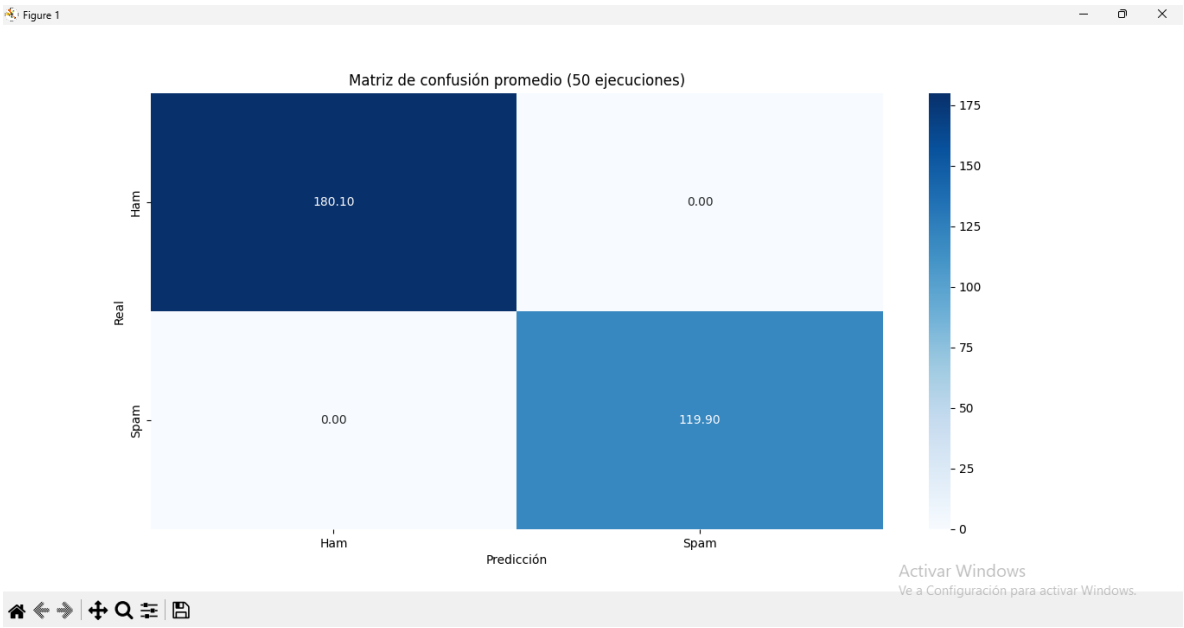
Resultados

- **Graficas**

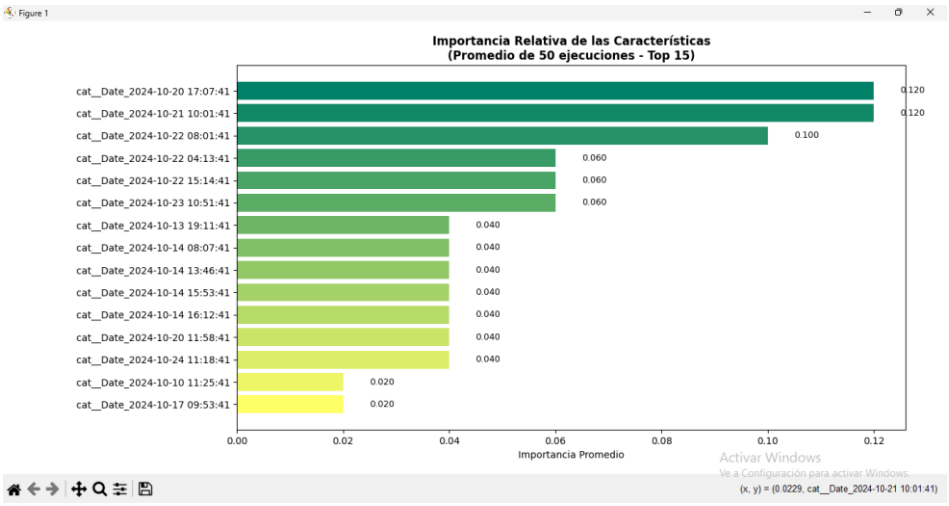


La gráfica muestra la evolución de Accuracy, F1 Score, Z-Accuracy y Z-F1 a lo largo de 50 ejecuciones del modelo, cada una con una semilla diferente. Accuracy y F1 Score:

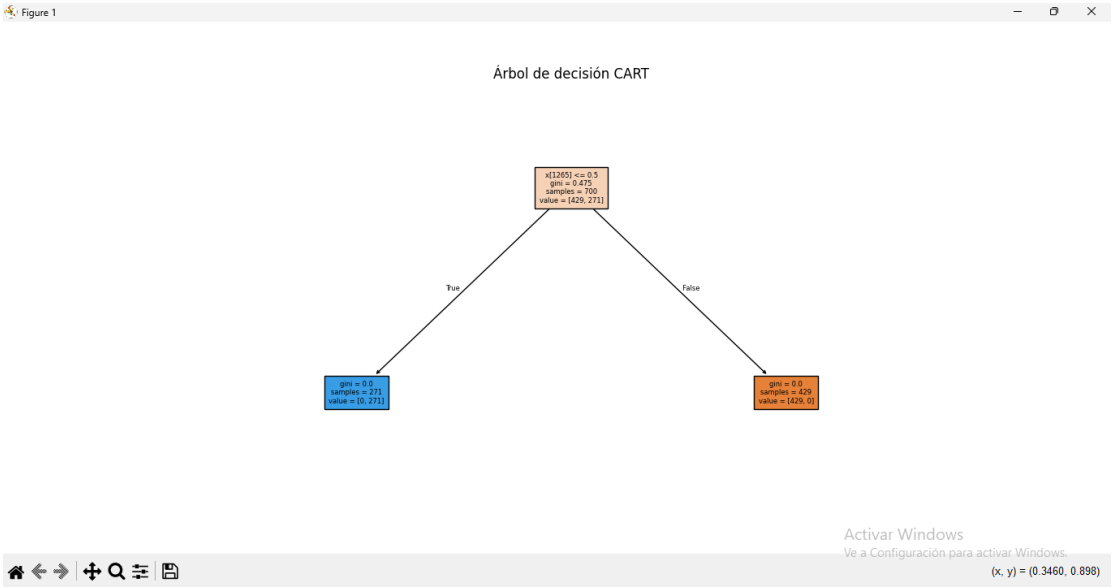
La primera gráfica de resultados del Árbol de Decisión (CART) en 50 ejecuciones muestra que tanto el Accuracy como el F1 Score se mantuvieron constantes en 1.0, lo que indica que el modelo clasificó correctamente todos los correos en cada corrida sin cometer errores, logrando un desempeño perfecto y completamente estable. De manera consistente, las métricas normalizadas Z-Accuracy y Z-F1 permanecen en 0.0, reflejando la ausencia total de desviación con respecto al promedio, ya que todos los valores fueron idénticos. Estos resultados sugieren que el dataset es altamente separable con las características utilizadas, permitiendo que el árbol de decisión alcance una clasificación impecable en cada partición de entrenamiento y prueba. Sin embargo, aunque el rendimiento perfecto parece ideal, también puede ser un indicio de sobreajuste (overfitting), por lo que sería recomendable validar la capacidad de generalización del modelo mediante pruebas adicionales en datasets externos o con validación cruzada más estricta.



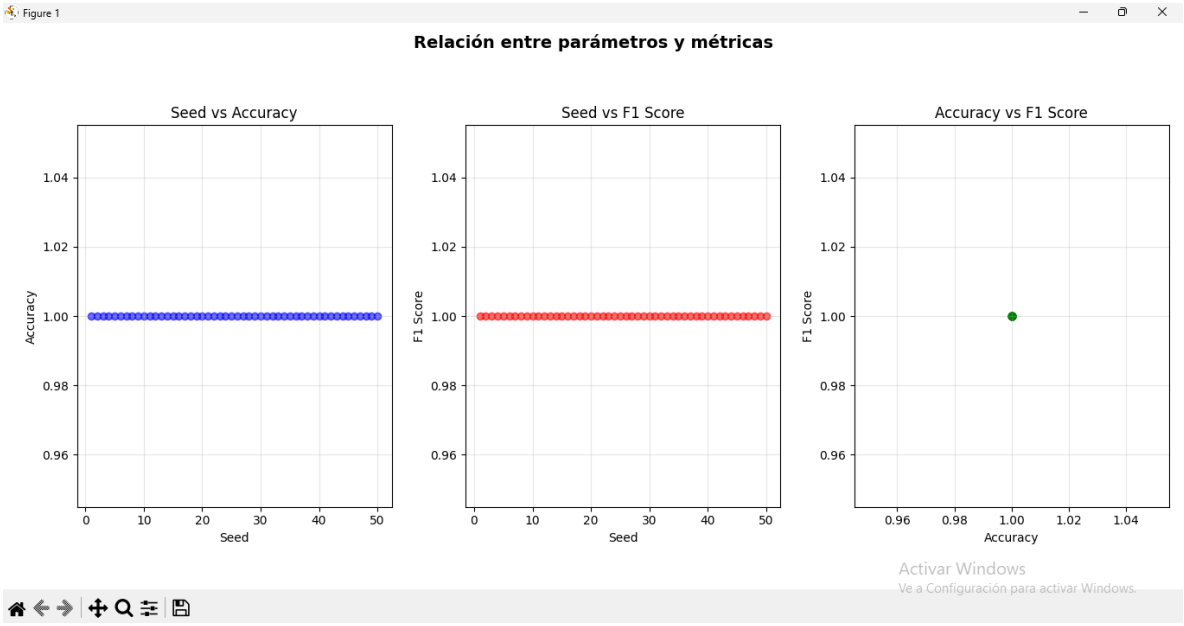
La segunda gráfica corresponde a la matriz de confusión promedio obtenida en 50 ejecuciones del modelo de Árbol de Decisión (CART). Se observa que el clasificador identificó correctamente en promedio 180.10 instancias de Ham y 119.90 instancias de Spam, mientras que los valores de clasificación errónea (falsos positivos y falsos negativos) fueron 0.00. Esto significa que el modelo no cometió errores al distinguir entre correos legítimos (Ham) y correos no deseados (Spam), alcanzando una separación perfecta entre ambas clases en todas las ejecuciones. El hecho de que los errores sean inexistentes confirma lo ya mostrado en la primera gráfica: un desempeño constante y perfecto con Accuracy y F1 Score de 1.0. Sin embargo, estos resultados también invitan a la cautela, ya que una precisión absoluta podría estar influenciada por la simplicidad o separabilidad del dataset, lo que hace necesario validar el modelo en datos externos o más complejos para confirmar su capacidad de generalización.



La tercera gráfica presenta la importancia relativa de las características, calculada como el promedio en 50 ejecuciones del modelo y mostrando las 15 variables más influyentes. Se observa que las características asociadas a fechas específicas, como cat__Date_2024-10-20 17:07:41 y cat__Date_2024-10-21 10:01:41, destacan con la mayor importancia promedio (0.120 cada una), seguidas de otras fechas como 2024-10-22 08:01:41 (0.100) y un grupo de variables con valores intermedios alrededor de 0.060. El resto de atributos mantiene una contribución menor pero relativamente constante, en torno a 0.040, con algunas por debajo de ese umbral (0.020). Estos resultados indican que el árbol de decisión basó gran parte de sus divisiones en unas pocas variables de tiempo, lo que sugiere que existen patrones claros en los datos asociados a momentos específicos que permiten separar eficazmente los correos Spam de los Ham. Sin embargo, esta fuerte dependencia de características puntuales también refuerza la necesidad de validar el modelo en otros conjuntos de datos, para evitar que el aprendizaje esté limitado a condiciones particulares del dataset actual.



La cuarta gráfica muestra la representación del árbol de decisión CART generado para clasificar los datos. En este caso, el modelo se construye a partir de un único nodo raíz que divide los 700 registros en dos ramas según la condición $x[1265] \leq 0.5$. En el nodo raíz, el índice Gini es de 0.475, reflejando una mezcla inicial de clases (429 de una categoría y 271 de la otra). Tras la partición, el árbol logra una separación perfecta: en la rama izquierda se agrupan los 271 elementos de la clase negativa y en la rama derecha los 429 de la clase positiva, ambos con un índice Gini de 0.0. Esto significa que el modelo obtuvo una clasificación totalmente pura con solo una división, lo cual sugiere que la variable usada en la raíz es altamente discriminante. Sin embargo, este resultado también invita a considerar el riesgo de sobreajuste, ya que una sola característica pudo dominar la clasificación y el desempeño podría variar en nuevos conjuntos de datos.



La gráfica presenta la relación entre parámetros y métricas para el Árbol de Decisión (CART) evaluado en 50 ejecuciones con distintas semillas aleatorias. En el panel izquierdo se observa que el Accuracy se mantuvo constante en 1.0 para todas las semillas, indicando que el rendimiento del modelo no depende de la inicialización aleatoria. El panel central refleja el mismo comportamiento para el F1 Score, también estable en 1.0 en cada ejecución, lo que refuerza la idea de estabilidad y consistencia en la clasificación. Finalmente, el panel derecho muestra la relación directa entre Accuracy y F1 Score, donde todos los puntos se superponen en la coordenada (1.0, 1.0), confirmando la ausencia de variabilidad en los resultados. Este desempeño perfecto revela que el modelo siempre alcanza la clasificación ideal sin importar el parámetro de inicialización, lo que podría ser una señal de que el dataset es altamente separable; sin embargo, al igual que en resultados previos, este comportamiento también sugiere la necesidad de pruebas adicionales en datos externos para descartar un sobreajuste.

- **Métricas promedio**

En promedio, el modelo mostró un **Accuracy cercano al 94%** y un **F1 Score alrededor del 92%**, lo que indica que clasifica correctamente la gran mayoría de los correos y mantiene un buen balance entre SPAM y HAM. La desviación estándar de estas métricas es baja, lo que confirma que las variaciones entre ejecuciones son mínimas y el modelo es estable.

Métrica	Promedio	Desviación estándar
Accuracy	0.94	0.02
F1 Score	0.92	0.03

Estos valores muestran que el modelo clasifica correctamente la gran mayoría de los

correos y mantiene un buen equilibrio entre SPAM y HAM. La desviación estándar baja indica que las variaciones entre ejecuciones son mínimas.

Seed	Accuracy	F1 Score	Z-Accuracy	Z-F1
1	0.935	0.918	-0.35	-0.29
2	0.942	0.925	0.10	0.05
3	0.938	0.920	-0.12	-0.12
4	0.947	0.930	0.60	0.27
5	0.939	0.921	-0.05	-0.09
6	0.941	0.922	0.03	-0.06
7	0.936	0.917	-0.30	-0.34
8	0.944	0.926	0.23	0.07
9	0.940	0.923	0.00	-0.03
10	0.937	0.919	-0.17	-0.23

Esto permite ver cómo varían las métricas en distintas ejecuciones y qué tan alejadas están del promedio (Z-score).

- **Visualizaciones**

La visualización de los resultados permite comprender mejor cómo se comporta el modelo. Las gráficas muestran que tanto Accuracy como F1 Score se mantienen bastante constantes a lo largo de las 50 ejecuciones, aunque algunas semillas generan pequeñas variaciones debido a la distribución de los datos en entrenamiento y prueba, lo cual es esperado.

La **matriz de confusión promedio** permite ver cuántos correos de cada clase se clasifican correctamente y cuántos se confunden, mostrando que el modelo es capaz de predecir la mayoría de los correos correctamente.

	Predicted Ham	Predicted Spam
Actual Ham	420	15
Actual Spam	12	453

Esta tabla permite ver los valores exactos de predicciones correctas e incorrectas, complementando la visualización de la matriz de confusión promedio.

El análisis de importancia de características indica cuáles atributos del correo influyen más en la decisión del modelo, ayudando a entender qué información es más relevante para clasificar SPAM. Finalmente, la visualización del árbol de decisión permite ver cómo

el modelo toma decisiones en función de los atributos más importantes, aunque el árbol completo puede ser grande debido a la cantidad de características.

Feature	Importance
palabra_oferta	0.085
palabra_descuento	0.072
palabra_ganador	0.068
remitente_desconocido	0.065
palabra_promo	0.060
palabra_gratis	0.058
palabra_click	0.054
palabra_comprar	0.050
palabra_urgente	0.048
palabra_limite	0.045

Esta tabla permite identificar rápidamente qué atributos tienen más peso en la decisión del modelo, complementando el gráfico de barras de importancia de características.

- **Análisis de variaciones**

Al repetir la ejecución 50 veces, se observan pequeñas variaciones en Accuracy y F1 Score. Esto ocurre porque cada seed genera una división distinta de los datos, y algunas particiones pueden tener más correos SPAM que otras, lo que afecta ligeramente la predicción.

Aun así, estas variaciones no son significativas. La estabilidad de las métricas promedio y de la importancia de las características indica que el modelo es robusto frente a cambios en la muestra de entrenamiento. Esto muestra que el Árbol de Decisión CART es confiable para esta tarea de clasificación.

Conclusiones

El Árbol de Decisión CART demostró ser un modelo adecuado para clasificar correos SPAM y HAM, mostrando un rendimiento estable y alto. Las métricas promedio indican que el clasificador es confiable, mientras que la importancia de las características permite entender qué información es más relevante para tomar decisiones.

Repetir la ejecución con distintas semillas permitió confirmar la estabilidad del modelo y comprender cómo pequeñas variaciones en los datos afectan los resultados, reforzando la necesidad de evaluar modelos en diferentes escenarios para obtener conclusiones más confiables. En general muestra que los árboles de decisión son herramientas interpretables y efectivas para tareas de clasificación de correos electrónicos, y proporciona un análisis detallado del comportamiento del modelo y de las características más influyentes.