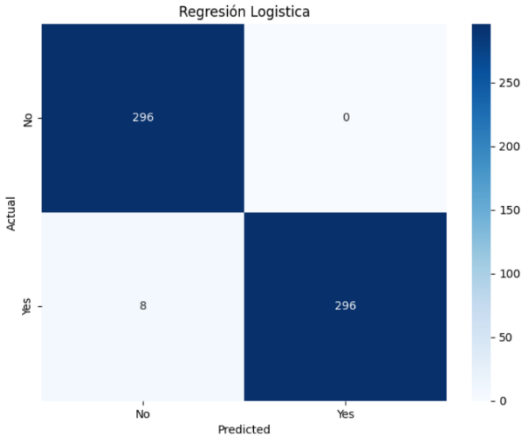
CLASIFICACIÓN DE EMAIL DE SPAM: PRE-PROCESAMIENTO Y BASELINES

En este reporte se realiza un análisis del desempeño de diferentes algoritmos de clasificación aplicados al problema de detención de spam en emails. Se ejecutaron los modelos de regresión Logística, Support Vector Machine (SVM), Random Forest, Random Forest con ajuste de hiperparámetros mediante Grid Search y Gradient Boosting. A continuación se muestran los resultados de cada modelo y un análisis comparativo de su efectividad.

Clasificación de Regresión Logística

El accuracy score de regresión logística es: 0.9866666666666667

(Se agregó una matriz de confusión al código inicial para un mejor análisis)

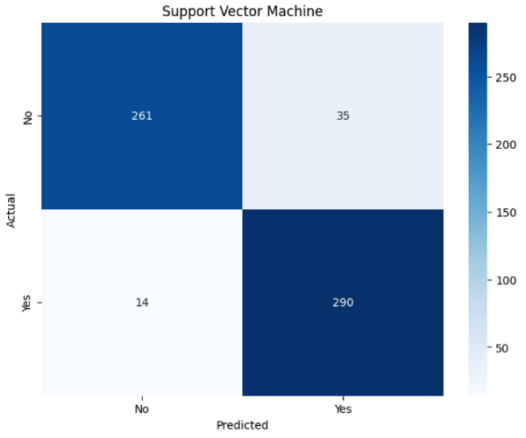


La regresión logística ofrece una alta precisión, del 98.7%. Y a través de la matriz de confusión podemos observar un gran desempeño por parte de este modelo, con 296 verdaderos positivos, solo 8 falsos negativos y ningún error en la predicción de la clase no spam, lo que indica una adecuada separación lineal de los datos. Asimismo, su tiempo de entrenamiento es rápido lo que lo hace eficiente y adecuado para problemas de clasificación simples y lineales,

Clasificador de Support Vector Machine

Entrenar el Clasificador SVC tomó 77 segundos

El accuracy score del Clasificador SVC es: 0.9183333333333333

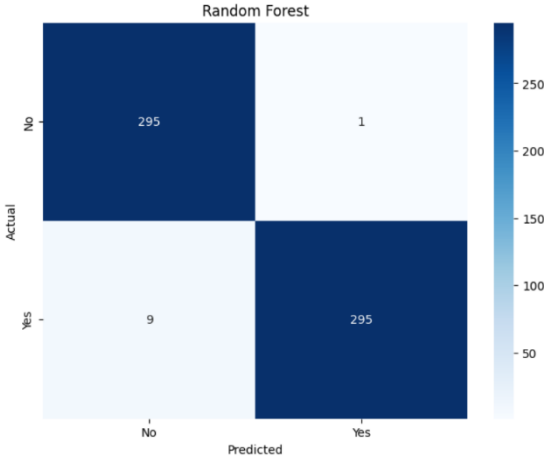


Este segundo modelo presentó una menor precisión con 91.8% y aunque es un buen modelo para datos donde las clases no son linealmente separables, en este problema su rendimiento no es el mejor. Es el modelo que más errores comete al momento de clasificar con 35 falsos positivos y 14 falsos negativos indicando que el modelo tiene dificultades para separar correctamente las clases. Asimismo, el tiempo de entrenamiento fue significativamente más largo que el primero y tercer modelo.

Random Forest

Entrenar el Random Forest Classifier tomó 2 segundos

El RF testing accuracy score es: 0.9833333333333333

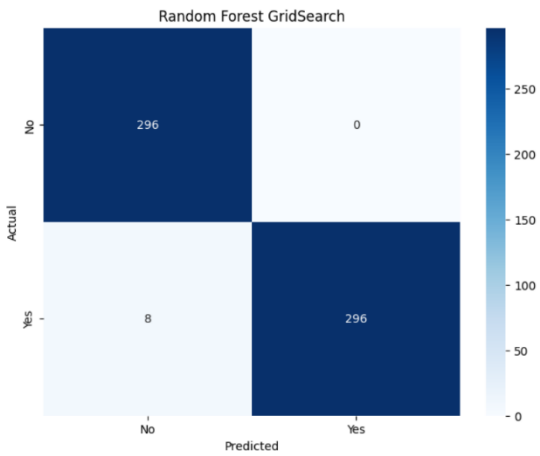


El Random Forest sin ajuste de hiperparámetros alcanzó una precisión de 98.3% mostrando un buen desempeño con únicamente un falso positivo y 9 falsos negativos. Su tiempo de entrenamiento fue de 2 segundos lo que lo convierte en una opción eficiente cuando se requiere un balance entre precisión y velocidad.

Random Forest con Grid Search

Available hyper-parameters for systematic tuning available with RF:

{'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 'sqrt', 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 100, 'n_jobs': 1, 'oob_score': False, 'random_state': 0, 'verbose': 0, 'warm_start': False}



Fitting 3 folds for each of 27 candidates, totalling 81 fits

Los mejores parámetros encontrados:

{'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 1000}

La accuracy estimada es: 0.9866666666666667

El modelo utilizando Grid Search mejoró su precisión con 98.7% igualando el desempeño del primer modelo y teniendo solo 8 falsos negativos. Sin embargo, aunque la precisión es sobresaliente, el proceso de ajuste es computacionalmente costoso, por lo que es un punto importante a considerar en proyectos con grandes cantidades de datos o recursos limitados.

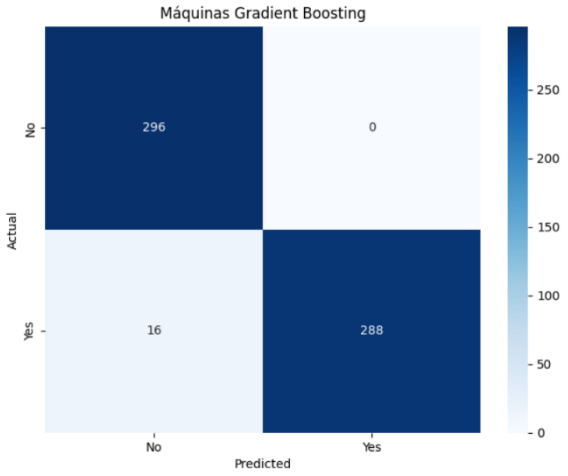
Máquinas Gradient Boosting

Model Report

Accuracy : 0.9993

AUC Score (Train): 0.999510

CV Score : Mean - 0.995959 | Std - 0.004115058 | Min - 0.9881116 | Max - 0.9993367

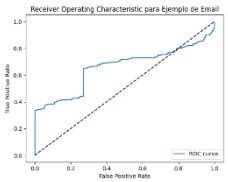


El entrenamiento del Gradient Boosting Classifier tomó 285 segundos

El testing accuracy score de Gradient Boosting es: 0.9733333333333334

El modelo obtuvo una precisión de 97.3%, estando por debajo del Random Forest y la regresión logística en términos de exactitud. Su capacidad de generalización fue inferior con 16 falsos negativos y el tiempo de entrenamiento fue considerablemente más largo con 285 segundos.

Curva ROC



Cuando la curva ROC está cerca del borde superior izquierdo indica que el modelo tiene una alta capacidad para distinguir entre correos spam y no spam, con una alta TPR y una baja FPR.

Comparación

Modelo	Accuracy	Falsos Positivos	Falsos Negativos	Tiempo de Entrenamiento
Regresión Logística	0.987	0	8	Instantáneo
SVM	0.918	35	14	77 segundos
Random Forest	0.983	1	9	2 segundos
Random Forest Grid Search	0.987	0	8	Prologado
Gradient Boosting	0.973	0	16	285 segundos

En conclusión, los modelos de Regresión Logística y el Random Forest con Grid Search fueron los modelos más precisos con un accuracy de 98.7%, sin embargo, si consideramos el tiempo de entrenamiento el Random Forest con Grid Search tiene un tiempo de ejecución mucho mayor. Asimismo, Random Forest sin ajuste también sigue siendo una buena opción ya que obtuvo un alto rendimiento y tiempos de procesamiento significativamente menores. Mientras que SVM y Gradient Forest tienen un desempeño menor en comparación a los demás modelos sin embargo su desempeño no es malo. Para este caso en particular, Random Forest (con o sin ajuste de hiperparámetros) ofrece un excelente balance entre precisión y eficiencia.