



# Universidad Autónoma de Nuevo León

## Facultad de Ciencias Físico Matemáticas

## Maestría en Ciencia de Datos

## Materia: Aprendizaje Automático

## PIA

## Modelo de Clasificación

Profesor: José Anastacio Hernández Saldaña

Alumno: Francisco David Treviño Bautista

Matrícula: 562795

Grupo: 1

Fecha de entrega: 26 de julio de 2024

## Objetivo

Se usará un conjunto de datos con valores numéricos, para hacer un modelo que predecirá el valor de un campo denominado “engagement”.

## Conjunto de Datos

El conjunto de datos son archivos en formato .csv, que nos compartió el maestro para el proyecto:

- Train.csv
- Test.csv

Las variables independientes que integran ambos archivos son las siguientes, solamente no comparte el archivo de entrenamiento con el archivo de pruebas el campo a predecir (“engagement”):

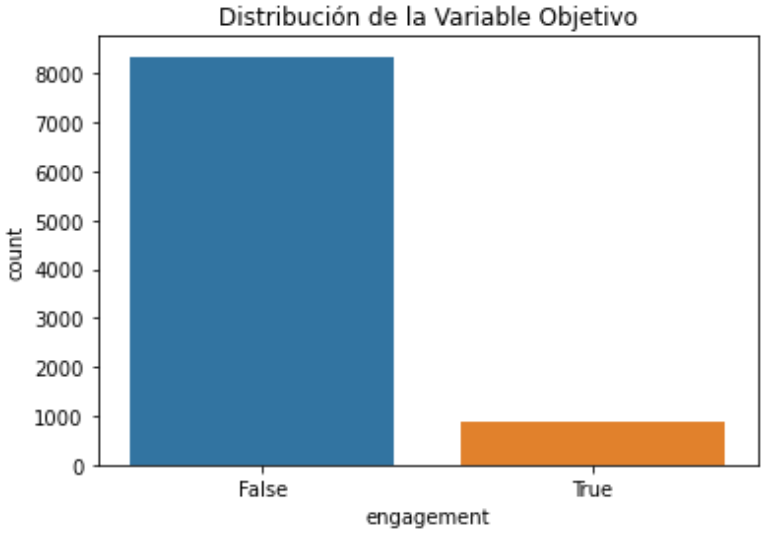
- id: identificador numérico entero consecutivo, sin aparente relación con la clasificación
- title_word_count: Cantidad de palabras en el título.
- document_entropy: Entropía del documento.
- Freshness: Frescura del contenido.
- Easiness_ Facilidad del contenido.
- fraction_stopword_presence: Fracción de palabras vacías.
- normalization_rate: Tasa de normalización.
- speaker_speed: Velocidad del hablante.
- silent_period_rate: Tasa de periodos silenciosos.
- engagement: Indicador de si hubo compromiso (True/False).

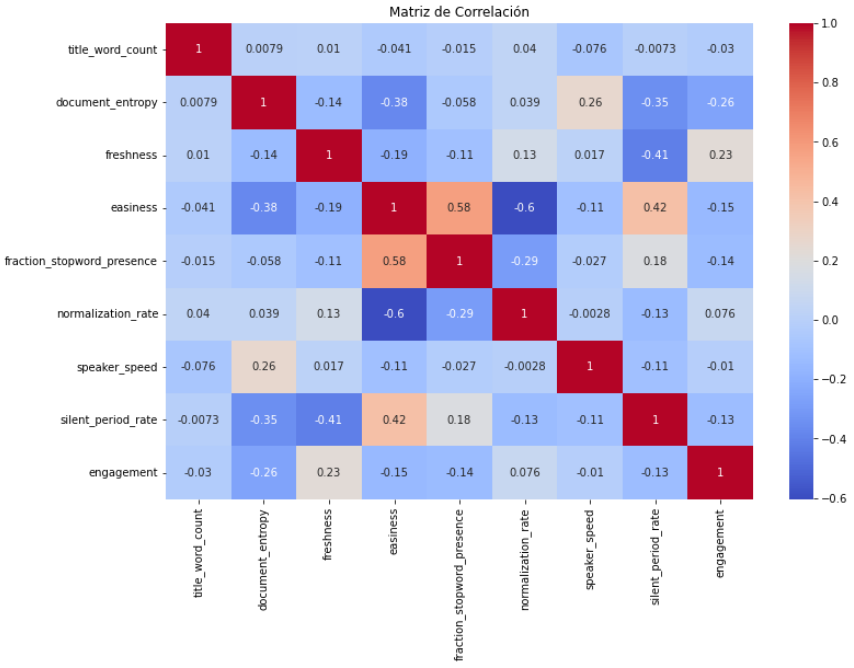
Se decide eliminar el campo “id” del conjunto de datos para el modelo, porque es un valor numérico entero consecutivo sin relación aparente.

# Análisis Exploratorio

El conjunto de datos tiene 9239 entradas. Los valores medios y desviaciones estándar varían significativamente entre las variables.

Se agrega unas gráficas para un análisis exploratorio con la idea de entender mejor la información del conjunto de datos y la relación entre las variables.



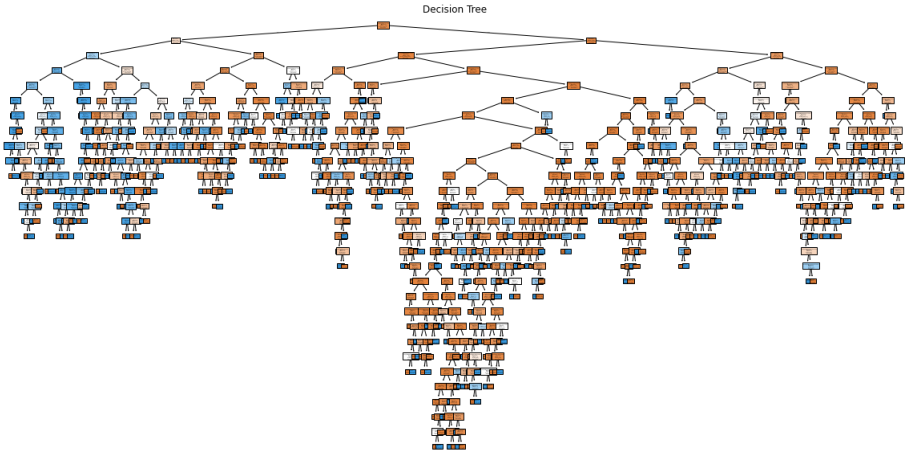



# Modelos de Clasificación

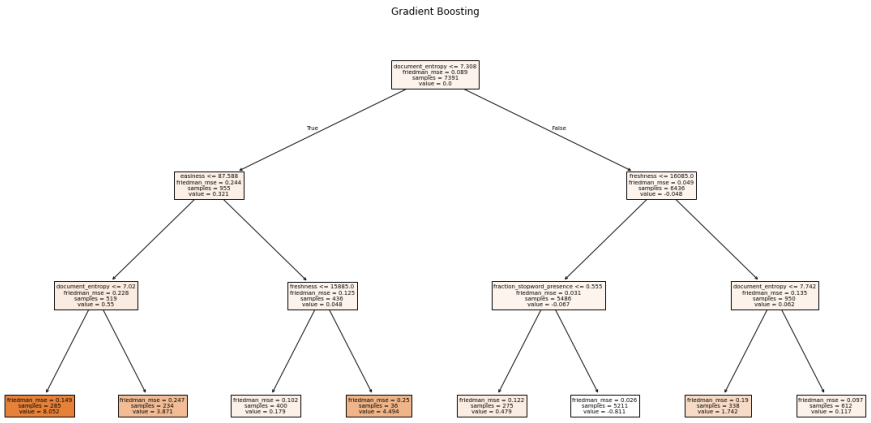
Se desarrolló código adicional para realizar las pruebas del conjunto de variables de clasificación, así como las variables objetivo, se evaluaron los siguientes modelos:

- Logistic Regression
- Support Vector Machine
- K-Nearest Neighbors
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- Neural Network
- XGBoost

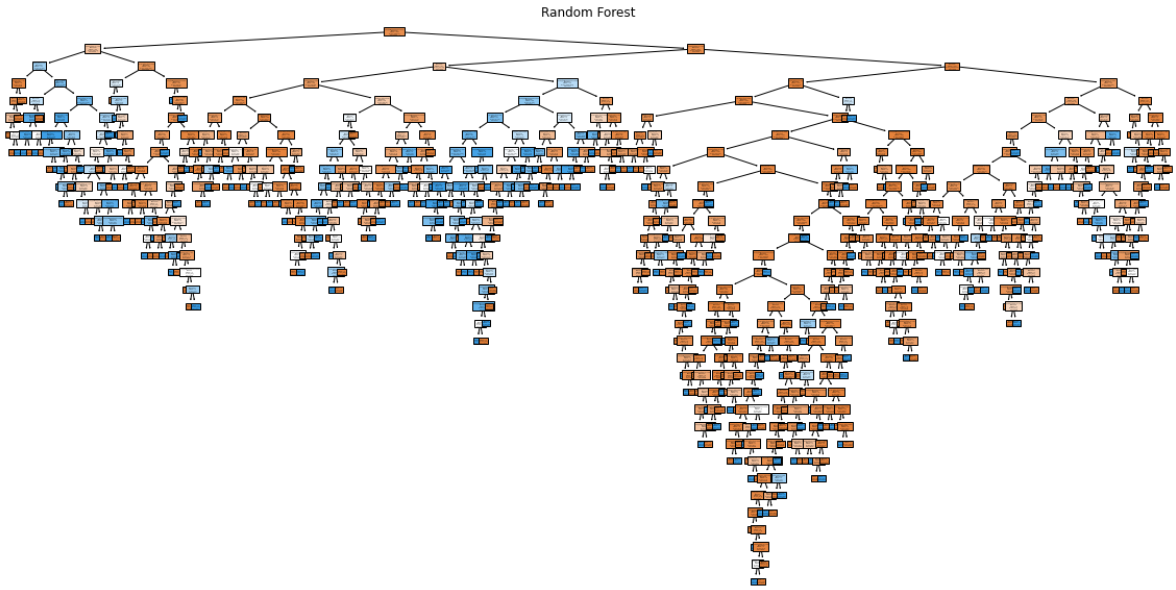
**Árbol Del Modelo Decision Tree Classifier**




**Árbol Del Modelo Gradient Boosting**



**Árbol Del Modelo Random Forest**



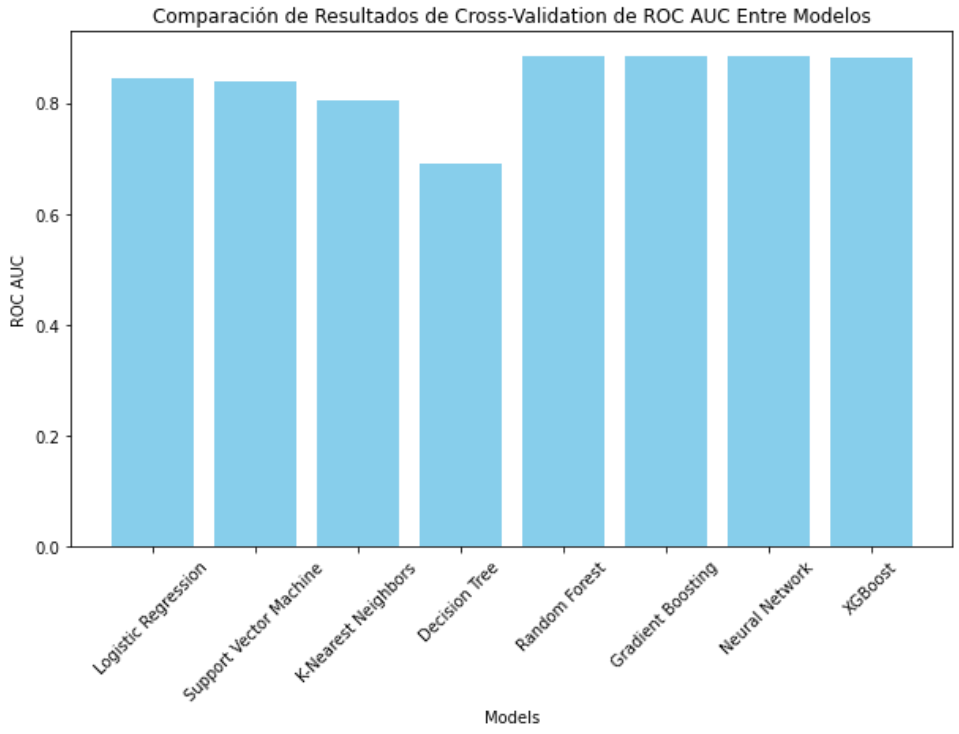

## Criterio de Selección del Modelo

En base a los dos conjuntos de datos de datos, uno de entrenamiento y otro de prueba, se deberá encontrar un modelo de clasificación con sus parámetros, utilizando cross-validation con el criterio ROC_AUC que resulte en un valor mayor de 0.75 en el conjunto de validación y prueba.

## Evaluación de Modelos

Se evaluaron varios modelos de clasificación usando validación cruzada (ROC AUC), seguido de un ajuste de hiperparámetros utilizando RandomizedSearchCV y GridSearchCV. Aquí se presentan los resultados para cada modelo:

| Modelo | Cross-Validation ROC AUC | Validation Accuracy | Precision | Recall | Validation ROC AUC | F1-Score |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.8446 | 0.9199 | 0.6049 | 0.297 | 0.8614 | 0.3984 |
| Support Vector Machine | 0.8376 | 0.9367 | 0.7727 | 0.4121 | 0.8544 | 0.5375 |
| K-Nearest Neighbors | 0.8033 | 0.928 | 0.66 | 0.4 | 0.81 | 0.4981 |
| Decision Tree | 0.6896 | 0.9004 | 0.4497 | 0.5152 | 0.7267 | 0.4802 |
| **Random Forest** | **0.8847** | **0.9334** | **0.6721** | **0.497** | **0.8983** | **0.5714** |
| Gradient Boosting | 0.8833 | 0.9356 | 0.6949 | 0.497 | 0.8915 | 0.5795 |
| Neural Network | 0.8841 | 0.9345 | 0.6667 | 0.5333 | 0.8942 | 0.5926 |
| XGBoost | 0.8826 | 0.9291 | 0.6197 | 0.5333 | 0.8988 | 0.5733 |

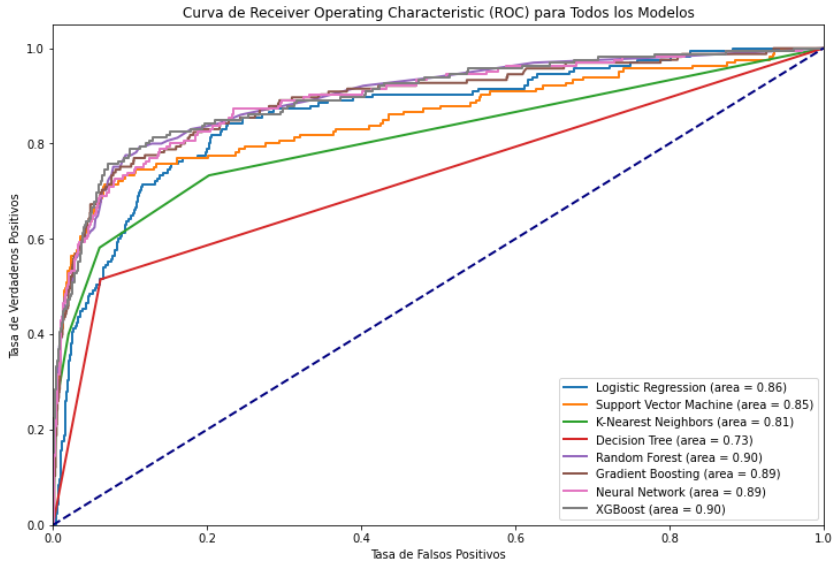


## Selección del Modelo

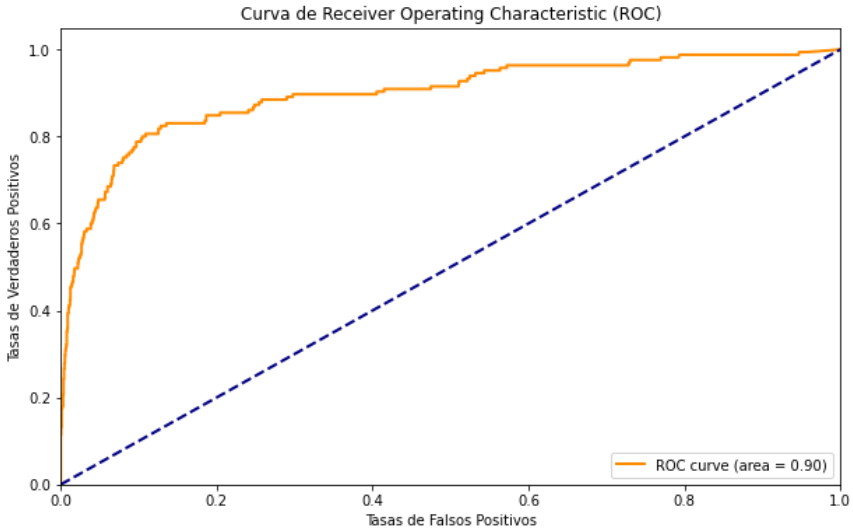
El mejor modelo en términos de la métrica Cross-Validation  ROC AUC fue Random Forest, con los siguientes mejores parámetros después de RandomizedSearchCV y GridSearchCV:

- RandomizedSearchCV: {'max_depth': 30, 'min_samples_split': 9, 'n_estimators': 84}
- GridSearchCV: {'max_depth': 30, 'min_samples_split': 9, 'n_estimators': 134}

La siguiente gráfica contiene la curva de ROC para todos los modelos evaluados, al final del eje X todas las curvas llegan a valores muy similares, y no se alcanza a distinguir.




Por lo anterior, se generó por separado una gráfica de la curva de ROC  para el mejor modelo, Random Forest



## Conclusiones

El modelo Random Forest mostró el mejor rendimiento global en Cross-Validation métrica ROC AUC. Los resultados muestran que este modelo tiene una mejor capacidad para distinguir entre clases en comparación con otros modelos evaluados.

El ajuste de hiperparámetros a través de RandomizedSearchCV y GridSearchCV mejoró aún más su rendimiento, lo que sugiere que los parámetros seleccionados están optimizados para este conjunto de datos.

Los gráficos y las tablas presentadas proporcionan mayor claridad de cómo cada modelo se desempeñó y facilitan la comparación entre ellos.

La curva ROC para cada modelo permite visualizar la tasa de verdaderos positivos frente a la tasa de falsos positivos, lo que es muy importante para evaluar la efectividad de los modelos en problemas de clasificación binaria.