Modelo de clasificación utilizando el criterio ROC_auc

Introducción

La lectura, desde mi punto de vista, es la habilidad más importante que toda persona debe de posser. México lamentablemente es un país que no tiene entre sus habitantes la cultura de leer. Lo cual es terrible ya que incentivar la lectura podrîa fomentar otras habilidades como el autodidactismo y la capacidad de concentración, memoria, comprensión, análisis, explicación y/o recitación.

El presente documento explica una base de datos que se base en medir las cualidaddes que posee un lector al momento de recitar en público. Las variables que forman dicha base son las siguientes:

Title_word_count: número de palabra en el título.

Document_entropy: diversidad o desorden de la información presentada en la lectura recitada.

Freshness: ¿Qué tan reciente es el documento recitado?

Easiness: facilidad de lectura o comprensión del documento.

Fraction_stopword_presence: fracción de palabras de parada presentes en el documento.

Normalization_rate: tasa utilizada para promover y facilitar el intercambio internacional sobre recursos archivísticos/documentos.

Speaker_speed: velocidad del hablante.

Silent_period_rate:tasa de periodos de silencio.

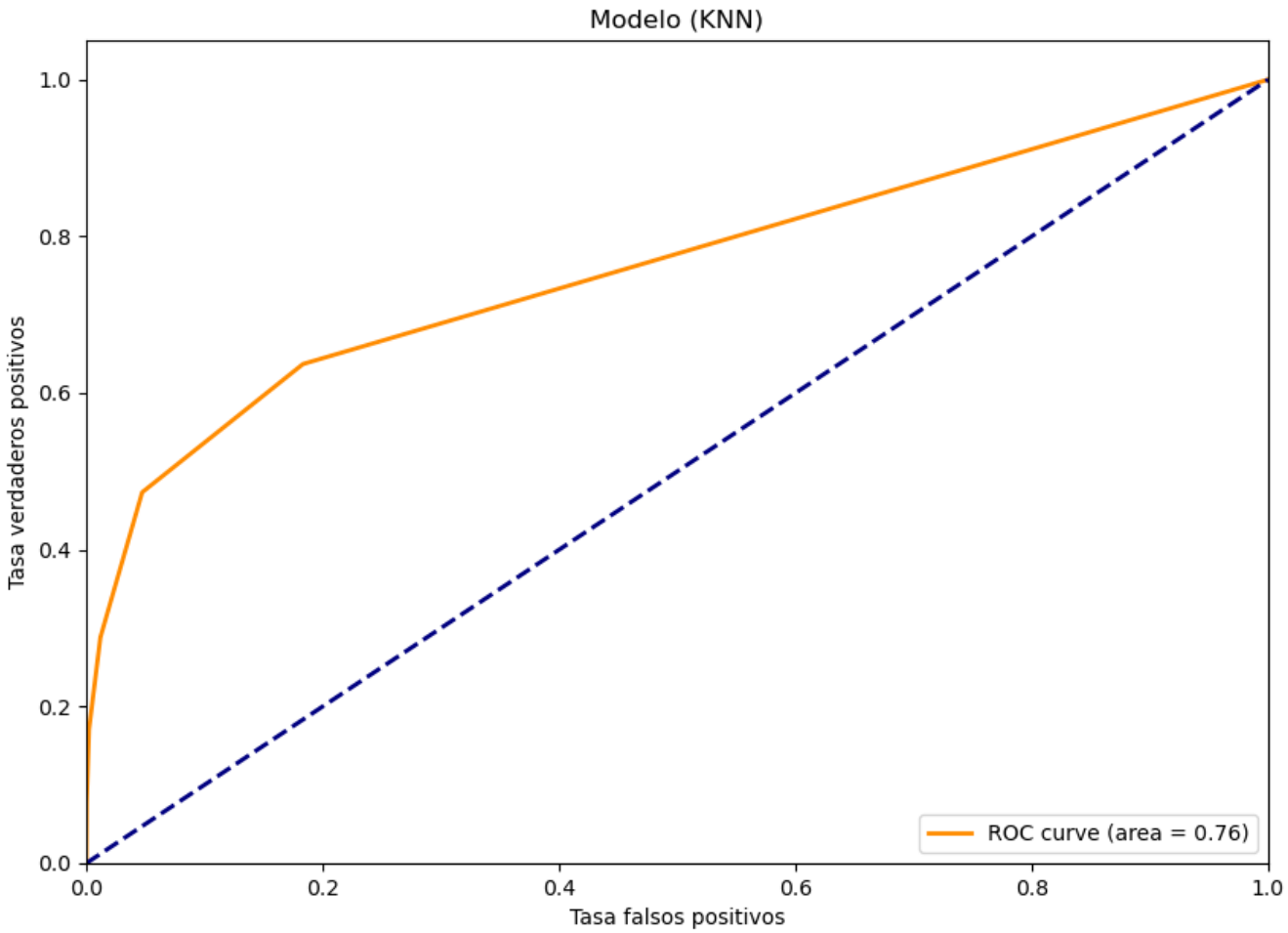
Engagement: nivel de compromiso o participación del lector con la audiencia.

De todas las variables consideradas, engagement es la que más nos interesa estudiar, ya que al ser una variable del tipo binaria, nos permitirá poder emplear modelos de clasificación mediante el criterio de ROC_auc. El criterio anterior, nos permitirá averiguar 2 cosas:

1. Determinar cuántos de nuestros datos, dependiendo si son falso o negativo en engagement, están registrados como verdaderos positivos y cuantos como falsos positivos.
2. Una vez que se decida la cantidad de verdaderos positivos y falsos positivos. Podremos determinar mediante el criterio ROC_auc la probabilidad de eficiencia con la que el modelo clasificó los resultados.

Modelos de clasificación

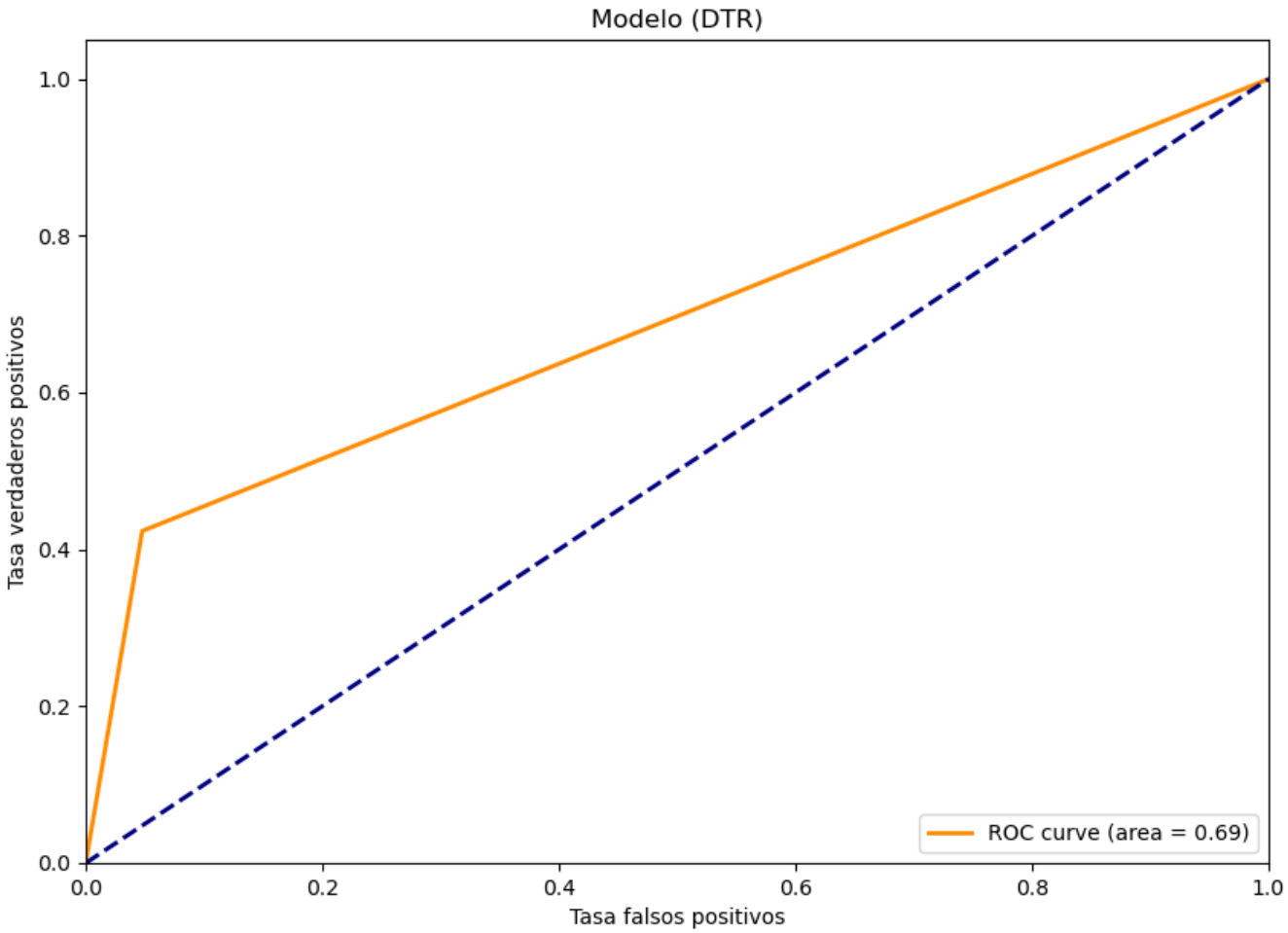
Modelo de clasificación KNN



Interpretación:

La eficiencia de este modelo para identificar a los datos verdaderos positivos y de los datos falsos positivos es de: 76%. Es un buen modelo.

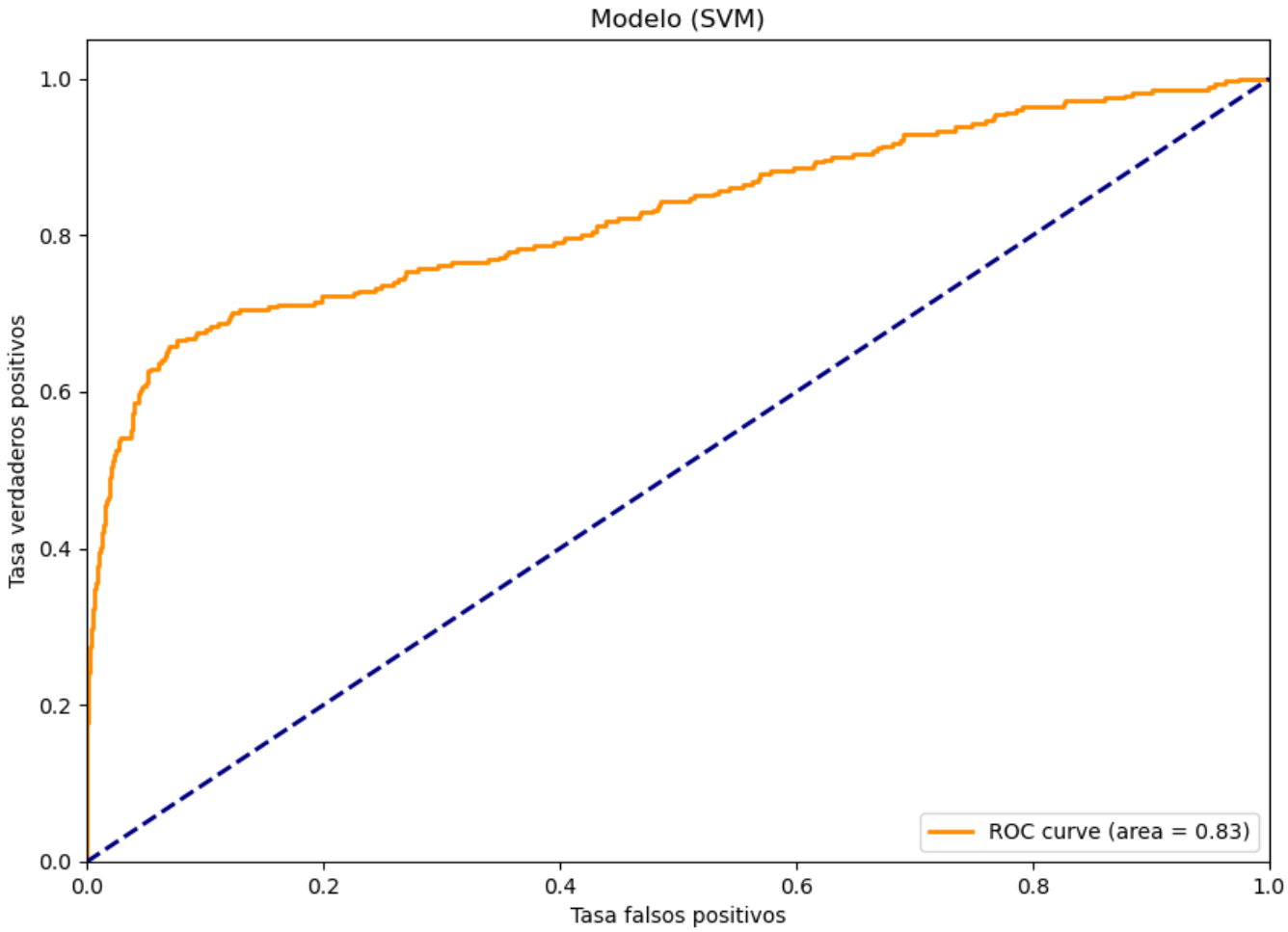
Modelo de clasificación árbol de decisión



Interpretación:

La eficiencia de este modelo para identificar a los datos verdaderos positivos de los datos falsos positivos es de: 69%. Es un modelo regular.

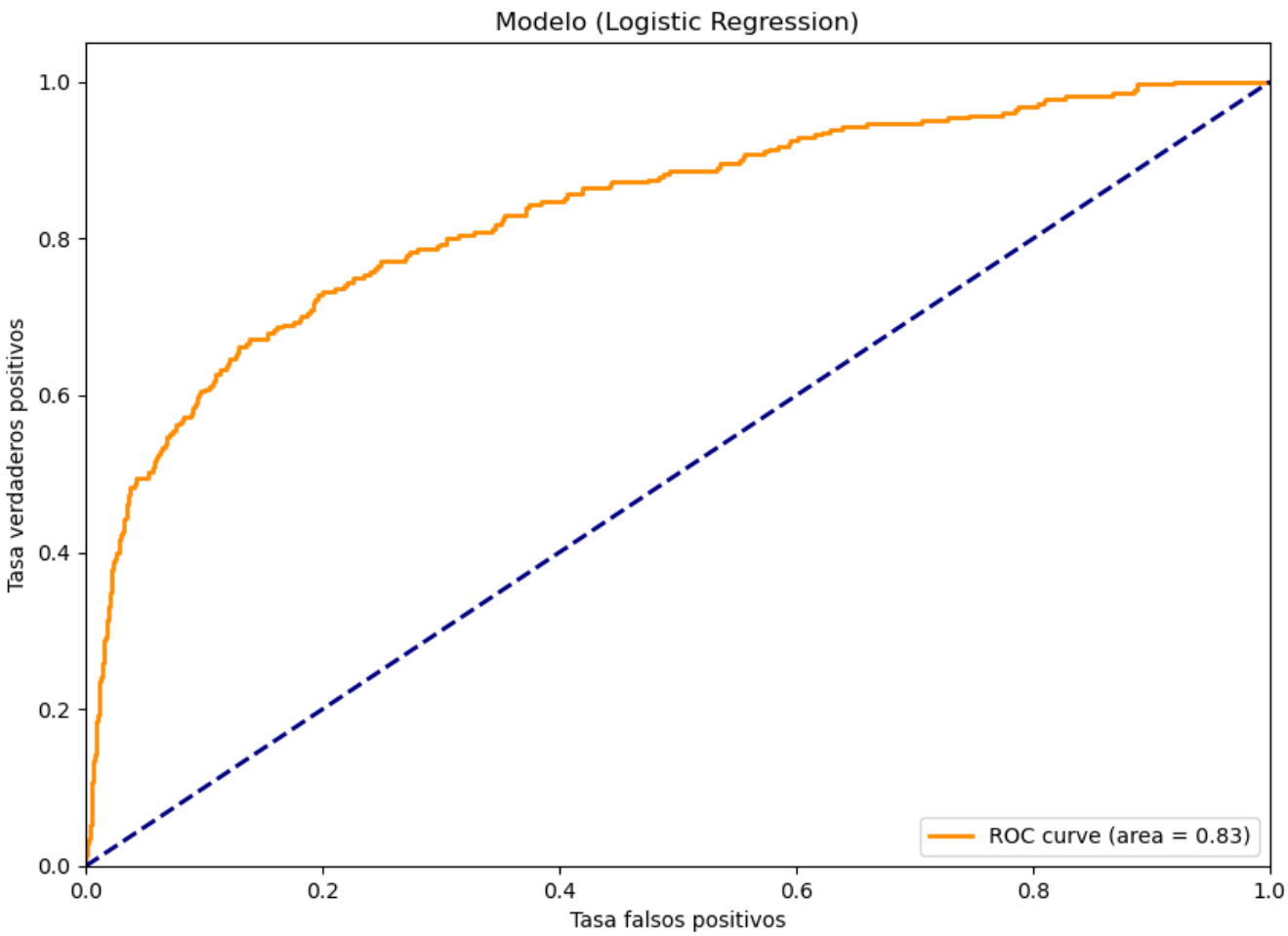
Modelo de clasificación SVM



Interpretación:

La eficiencia de este modelo para identificar a los datos verdaderos positivos de los datos falsos positivos es de: 83%. Es un modelo excelente.

Modelo de clasificación Logistic Regression



Interpretación:

La eficiencia de este modelo para identificar a los datos verdaderos positivos de los datos falsos positivos es de: 83%. Es un modelo excelente.

General

De acuerdo a los análisis realizados, los modelos con mejores porcentajes de eficiencia clasificación son SVM y Logistic Regression. Pero para poder decidir cual es el mejor de todos, debemos recurrir al método cross-validation.

Cross validation

En el análisis de cross-validation dividimos nuestro conjunto de datos en 5 secciones. Ya que la teoría indica que menos de 5 secciones proporciona menos variabilidad en los resultados, mientras que más de 5 proporciona más variabilidad en los resultados; 5 seciones se podría decir que es un punto medio entre ambos casos.

Puntos a revisar:

1. Cross-Validation ROC AUC Scores: puntaje obtenido en cada sección de cada modelo durante la aplicación de cross-validation.

2. Cross-Validation Mean ROC AUC Score: promedio de los puntajes Cross-Validation ROC AUC Scores en cada modelo.

3. Cross-Validation Std ROC AUC Score: desviación estándar baja de los puntajes Cross-Validation ROC AUC Scores en cada modelo

Impresión de resultados por modelo

KNN Cross-Validation ROC AUC Scores: [0.82026046 0.78685552 0.80017305 0.80583663 0.77807196]
KNN Cross-Validation Mean ROC AUC Score: 0.7982
KNN Cross-Validation Std ROC AUC Score: 0.0147

DTR Cross-Validation ROC AUC Scores: [0.70321743 0.70883293 0.68083454 0.71056499 0.70096828]
DTR Cross-Validation Mean ROC AUC Score: 0.7009
DTR Cross-Validation Std ROC AUC Score: 0.0106

SVM Cross-Validation ROC AUC Scores: [0.87708833 0.82083333 0.81760061 0.84863649 0.84466393]
SVM Cross-Validation Mean ROC AUC Score: 0.8418
SVM Cross-Validation Std ROC AUC Score: 0.0216

Logistic Regression Cross-Validation ROC AUC Scores: [0.85237144 0.82755129 0.85366074 0.85915696 0.83837399]
Logistic Regression Cross-Validation Mean ROC AUC Score: 0.8462
Logistic Regression Cross-Validation Std ROC AUC Score: 0.0116

Interpretación:

Si un modelo tiene un Cross-Validation Mean ROC AUC Score más alto que el resto, significa que es más eficiente para identificar a los datos verdaderos positivos de los datos falsos positivos. Si un modelo tiene un Cross-Validation Std ROC AUC Score más bajo que el resto, significa que tiene un rendimiento consistente en diferentes particiones del conjunto de datos.

Modelo con Cross-Validation Mean ROC AUC Score más alto

Logistic Regression Cross-Validation Mean ROC AUC Score: 0.8462
Logistic Regression Cross-Validation Std ROC AUC Score: 0.0116

Modelo con Cross-Validation Std ROC AUC Score más bajo

DTR Cross-Validation Mean ROC AUC Score: 0.7009
DTR Cross-Validation Std ROC AUC Score: 0.0106

Aunque su std no es el más bajo, solamente tiene 10 milesimas de diferencia respecto al modelo con un std más bajo. Por lo tanto, el mejor modelo de clasificación es el **modelo Logistic Regression**; con una eficiencia del 84.62% para identificar datos verdaderos positivos de los datos falsos positivos, y con un 1.16% de rendimiento inconsistente en diferentes particiones del conjunto de datos.