UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

PIA

Evaluación de mejor modelo de clasificación

Nombre: Armando Alexis Sepúlveda Cruz
Grupo: 003
Matricula: 1565746
Unidad de aprendizaje: Aprendizaje Automático
Profesor: JOSE ANASTACIO HERNANDEZ SALDAÑA

1. Objetivos

- Recopilar y organizar los conjuntos de datos adjuntos para su análisis y procesamiento.
- Realizar un análisis exploratorio de los datos para comprender sus características y posibles relaciones.
- Seleccionar y aplicar técnicas de preprocesamiento de datos, si es necesario, para mejorar la calidad de los datos y prepararlos para el modelado.
- Evaluar diferentes modelos de clasificación y seleccionar aquellos que sean más adecuados para el problema en cuestión.
- Ajustar los hiper parámetros de los modelos seleccionados utilizando técnicas de validación cruzada (cross validation) para optimizar su rendimiento.
- Evaluar el rendimiento de los modelos seleccionados utilizando el criterio ROC_auc en el conjunto de validación y prueba.
- Seleccionar el modelo con el mayor valor de ROC_auc en el conjunto de validación y prueba, asegurándose de que supere el umbral de 0.75.

2. Descripción de datos

Los datos contienen información sobre las características del texto y audio asociados a películas, así como la decisión de ver o no la película por parte de un usuario.

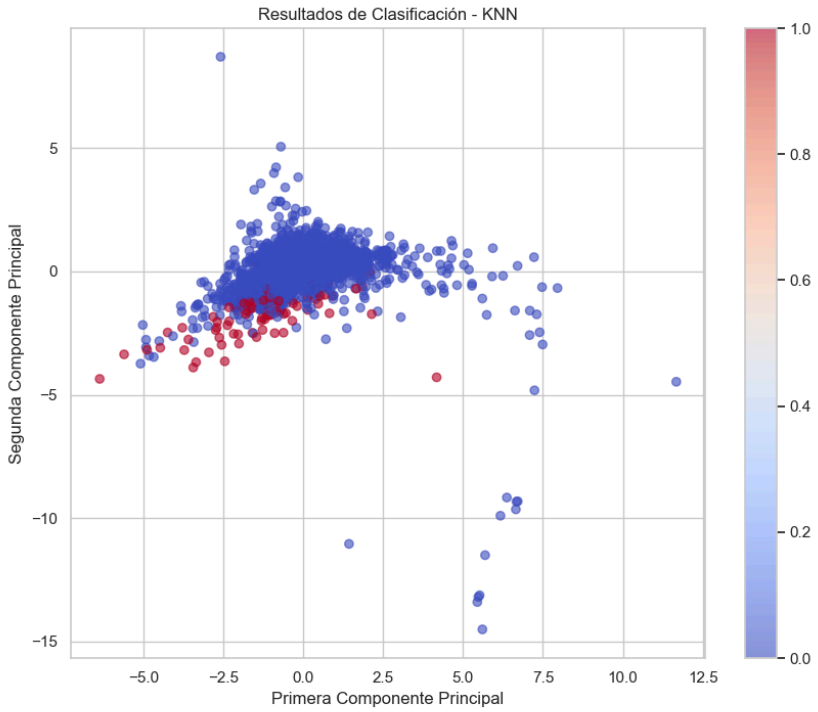
Variable	Tipo de dato	Descripción	Variable predictiva / respuesta	Valores posibles	Variable discreta o continua
title_word_count	int	Número de palabras en el título de la película	Predictiva	N/A	Discreta
document_entropy	float	Entropía del texto asociado a la película	Predictiva	[0, infinito)	Continua
freshness	int	Tiempo transcurrido desde el estreno de la película en días	Predictiva	[0, infinito)	Discreta

easiness	float	Legibilidad del texto asociado a la película	Predictiva	[0, 100]	Continua
fraction_stopword_presence	float	Fracción de palabras vacías en el texto asociado a la película	Predictiva	[0, 1]	Continua
normalization_rate	float	Tasa de normalización del texto asociado a la película	Predictiva	[0, 1]	Continua
speaker_speed	float	Velocidad promedio de habla en el audio asociado a la película	Predictiva	[0, infinito)	Continua
silent_period_rate	float	Fracción de tiempo de silencio en el audio asociado a la película	Predictiva	[0, 1]	Continua
engagement	boolean	Decisión de ver o no la película	Respuesta	True, False	Discreta

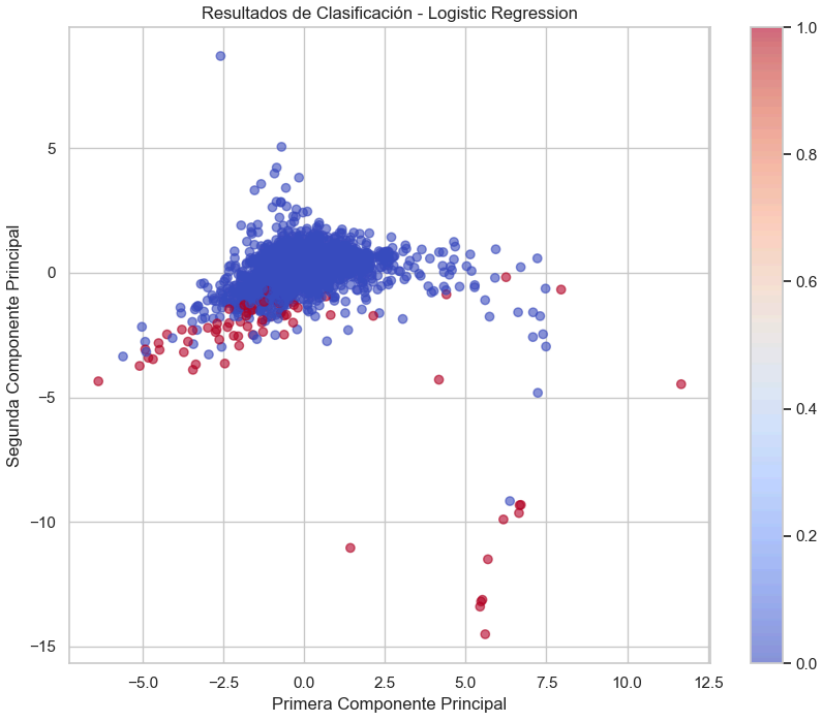
3. Gráficos de clasificación

Los gráficos se generando una reducción de dimensionalidad aplicando un análisis de componentes principales (PCA)

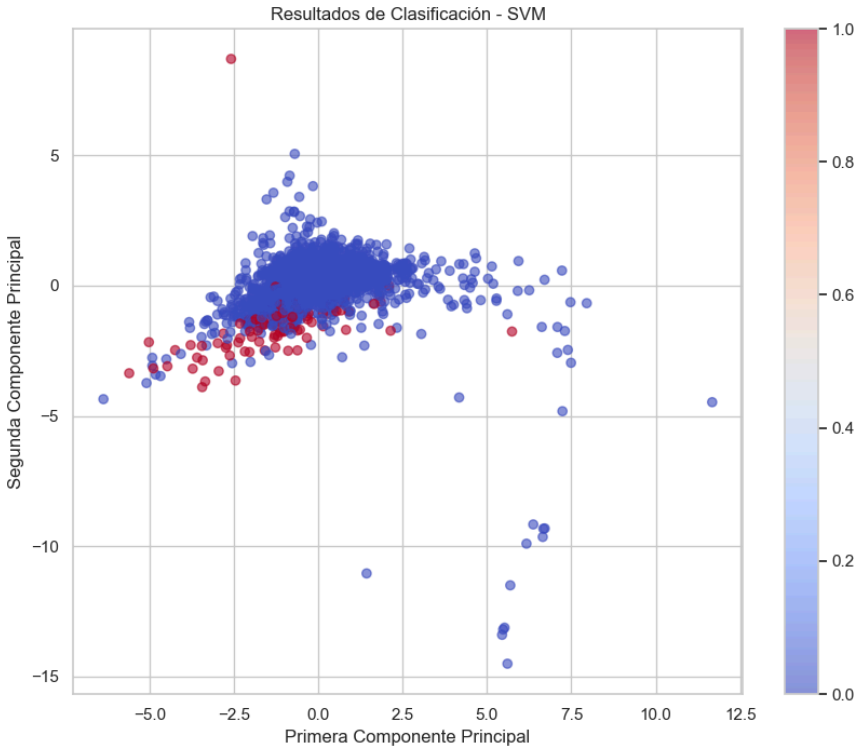
3.1 Clasificación KNN



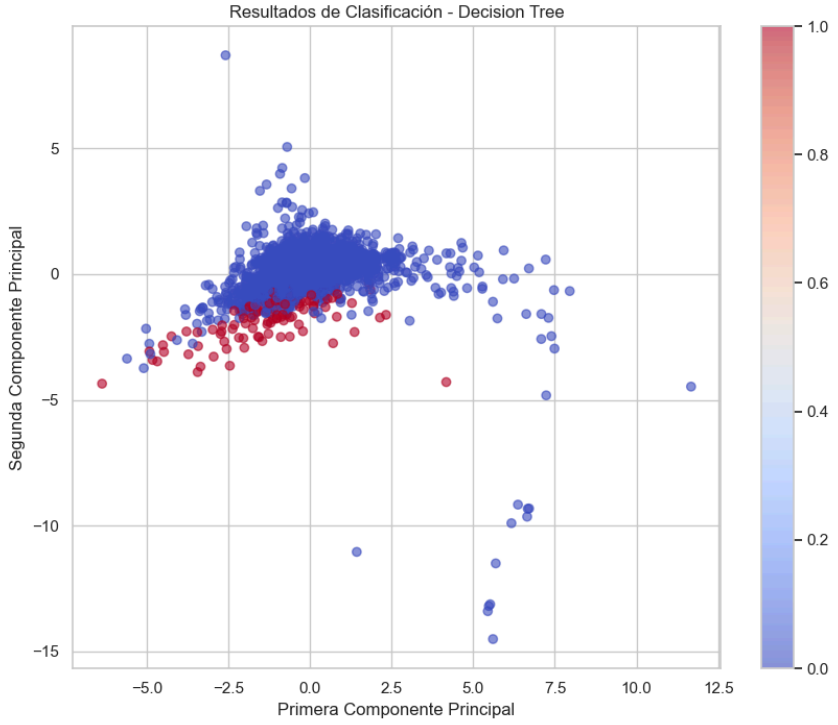
3.2 Clasificación de Regresión Logística



3.3 Clasificación Support Vector Machine



3.4 Clasificación Decision Tree

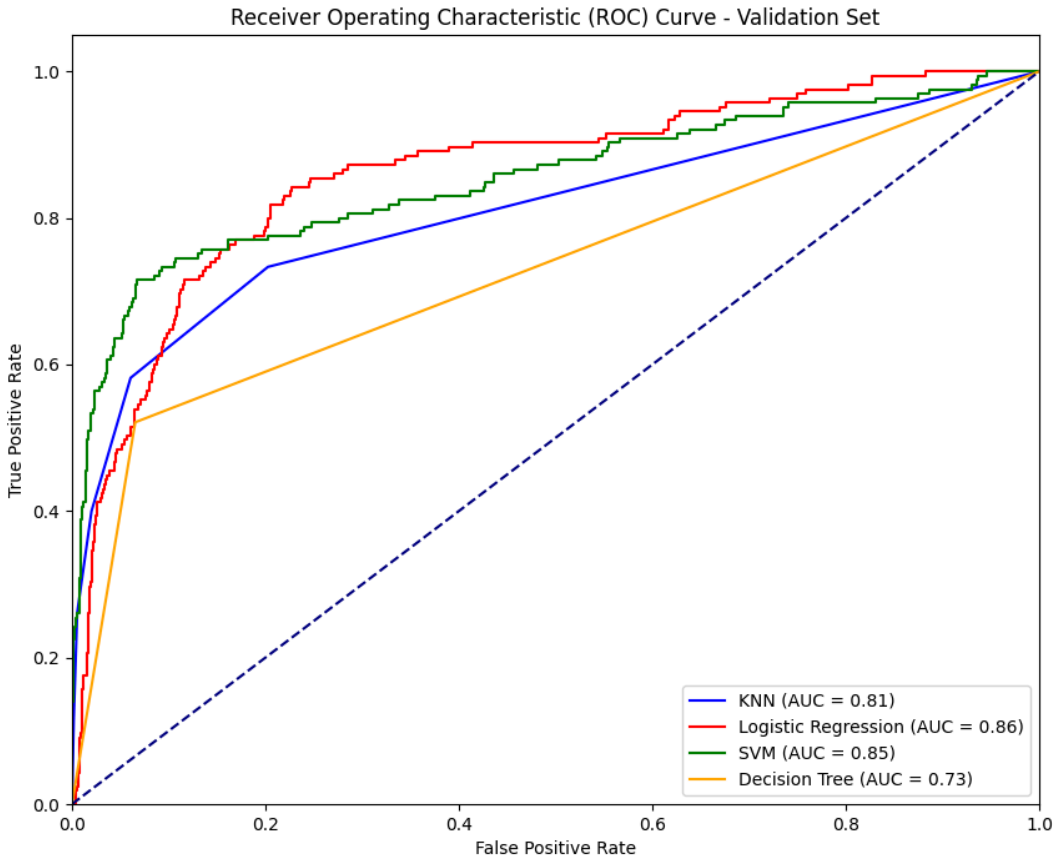


4. Evaluación

La curva ROC es un gráfico que muestra la relación entre la Tasa de Verdaderos Positivos (TPR) y la Tasa de Falsos Positivos (FPR) a medida que se varía el umbral de decisión del modelo de clasificación.

A continuación se muestran algunas gráficas comparativas entre cada uno de los modelos de clasificación sin usar grid search y utilizándolo.

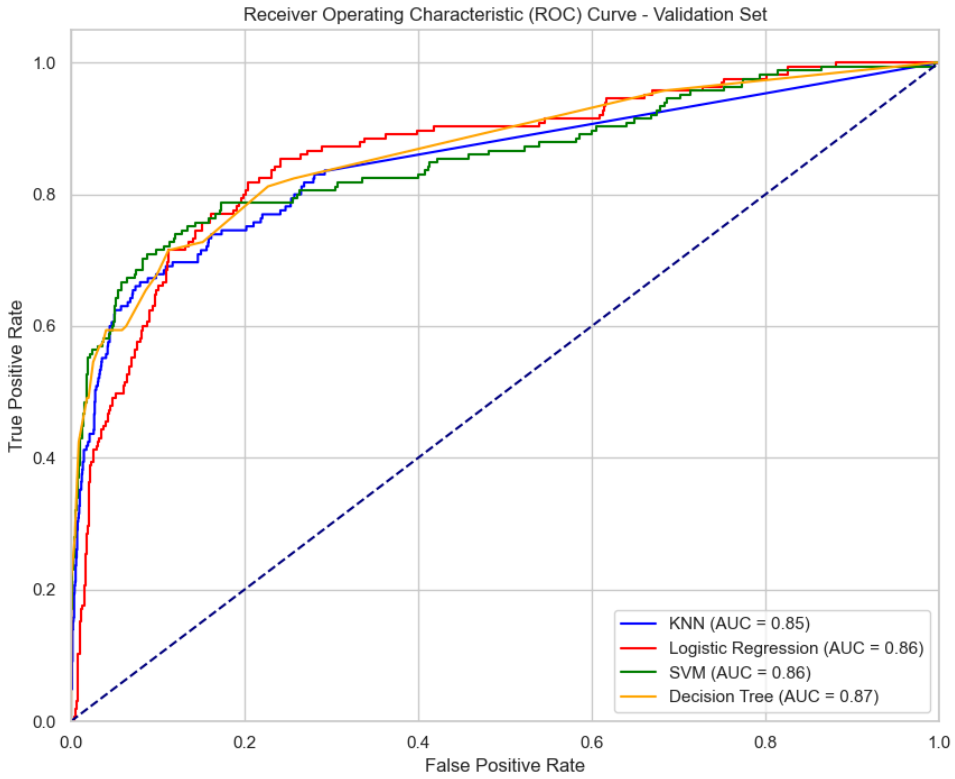
4.1 Roc Curve



Resultados de Clasificación sin Grid Search

	CV Score	Validation Score	Pos. Pred. Test	Mean Prob. Test
KNN	0.8045	0.8100	132	0.0896
Logistic Regression	0.8446	0.8614	92	0.1052
SVM	0.8375	0.8545	118	0.1093
Decision Tree	0.6965	0.7424	258	0.1117

4.2 Roc Curve con grid search



Resultados de Clasificación con Grid Search

	CV Score	Validation Score	Pos. Pred. Test	Mean Prob. Test
KNN	0.8371	0.8480	107	0.0882
Logistic Regression	0.8452	0.8622	87	0.1051
SVM	0.8436	0.8559	135	0.1058
Decision Tree	0.8405	0.8657	148	0.1050

4.3 Hiper parámetros

1. **KNN**:
    - Mejor número de vecinos (n_neighbors): 9
    - Mejor peso de distancia (weights): 'distance'
2. **Logistic Regression**:
    - Mejor valor de regularización (C): 0.1
    - Mejor penalización (penalty): 'l2'
    - Mejor método de optimización (solver): 'lbfgs'
3. **SVM**:
    - Mejor valor de regularización (C): 10
    - Mejor kernel: 'rbf'
4. **Decision Tree**:
    - Mejor profundidad máxima del árbol (max_depth): 5
    - Mejor número mínimo de muestras para dividir un nodo (min_samples_split): 2

5. Conclusiones

Considerando las puntuaciones de validación cruzada y validación, el modelo de **Decision Tree** parece tener el mejor desempeño general con la puntuación de validación más alta (0.8657). Sin embargo, Logistic Regression también muestra un buen desempeño con una puntuación de validación cruzada alta (0.8452) y una puntuación de validación cercana (0.8622).