Reporte de Evaluación de Modelos de Clasificación Binaria de Estrés.

1. Objetivo del Experimento

El objetivo del presente experimento fue identificar el modelo de aprendizaje automático más adecuado para la clasificación binaria de niveles de estrés en estudiantes, utilizando transcripciones de texto previamente normalizadas. La etiqueta objetivo fue la columna NivelPregunta, categorizada en "sí" (estrés) y "no" (sin estrés).

2. Preparación del Conjunto de Datos

Se utilizó un corpus etiquetado y previamente normalizado que incluye:

Conversión a minúsculas

Eliminación de tildes y caracteres especiales

Lematización con spaCy

Remoción de stopwords

El conjunto de datos fue dividido en 80% para entrenamiento y 20% para prueba mediante train_test_split con partición estratificada para conservar la proporción entre clases.

3. Representación del Texto (Vectorización)

Se exploraron distintas configuraciones de representación textual:

TF-IDF

Frecuencia

Binaria

Con combinaciones de n-gramas:

Unigramas

Bigramas

Trigramas

Uni+bi

Uni+tri

Cada combinación se guardó como un vectorizador .pkl para su uso posterior en la fase de modelado.

4. Modelos Evaluados

Se aplicaron los siguientes algoritmos de clasificación:

Logistic Regression

Support Vector Machine (SVC)

Multinomial Naïve Bayes

Random Forest

Para cada modelo se realizó una búsqueda en malla (GridSearchCV) con validación cruzada de 5 pliegues, optimizando la métrica f1_macro.

5. Resultados

Se evaluaron decenas de combinaciones de modelos y representaciones. El mejor rendimiento global fue obtenido por:

Modelo: SVC con kernel lineal

Vectorizador: TF-IDF con unigramas y bigramas

Hiperparámetros óptimos: C=1, kernel='linear'

F1 Macro (CV): 0.8093

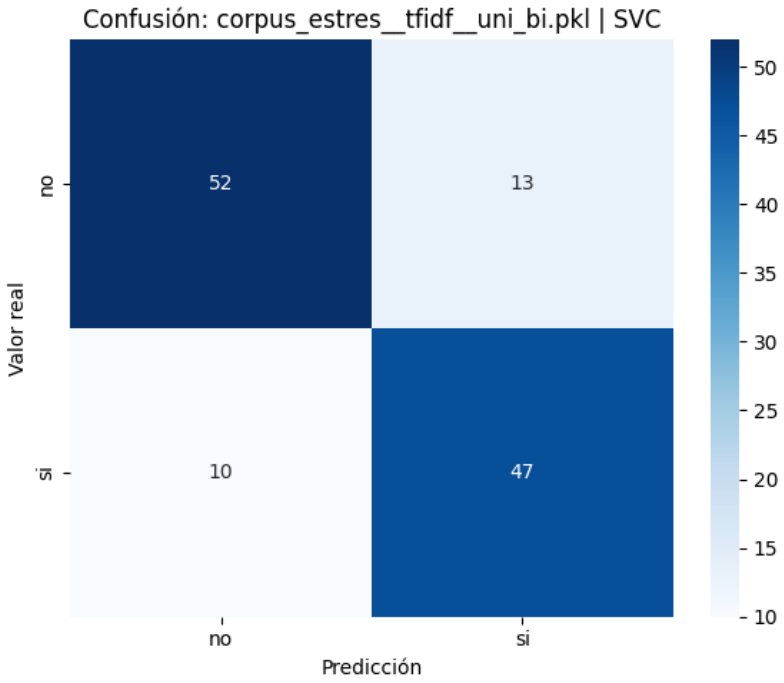
Precisión total (accuracy): 0.81

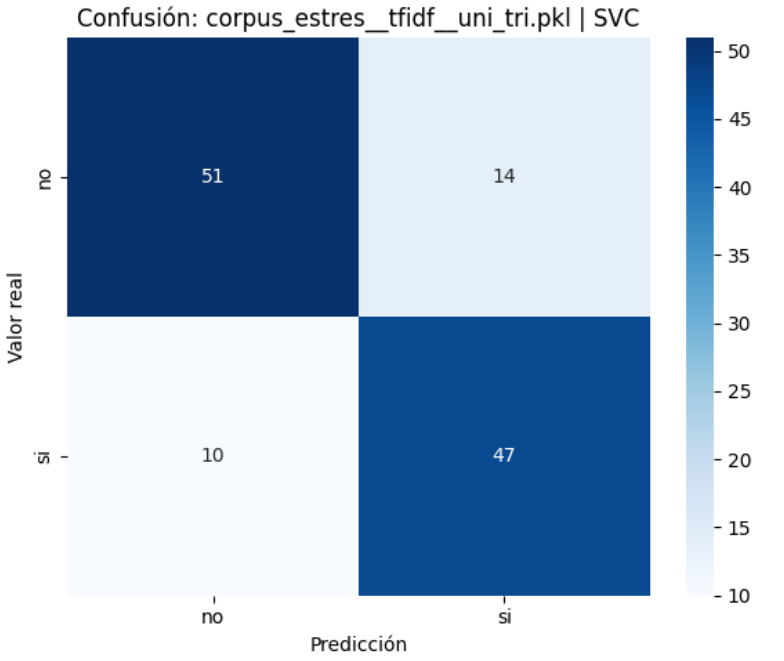
Además, otros modelos como MultinomialNB con TF-IDF uni+bi y Logistic Regression con TF-IDF uni+bi también alcanzaron puntuaciones superiores a 0.80 en F1 macro, lo que demuestra un rendimiento robusto en diferentes algoritmos bajo esta representación textual.

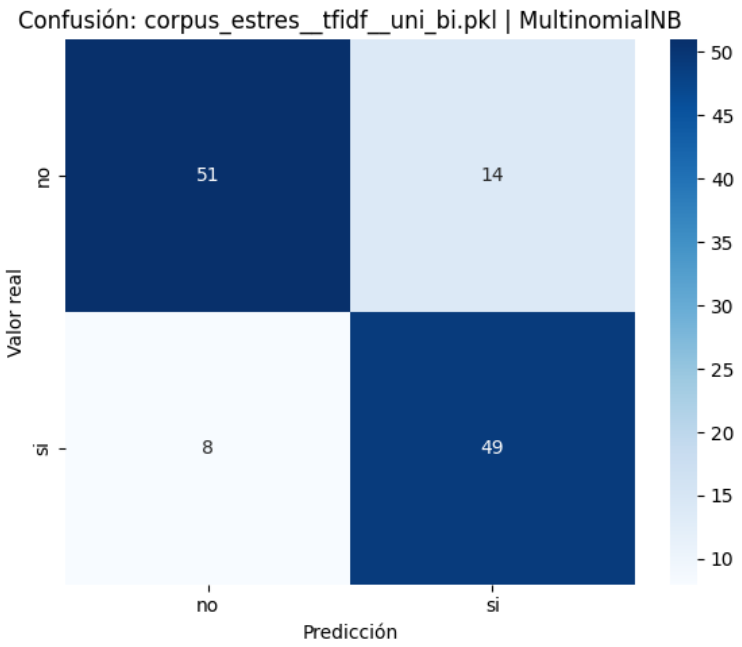
ML Method	ML Hyperparameters	Text Representation	Average F-Score Macro
SVC	{'C': 1, 'kernel': 'linear'}	corpus_estres__tfidf__uni_bi	0.8093
SVC	{'C': 10, 'kernel': 'linear'}	corpus_estres__tfidf__uni_tri	0.8072
MultinomialNB	{'alpha': 0.1}	corpus_estres__tfidf__uni_bi	0.8066
LogisticRegression	{'C': 10}	corpus_estres__tfidf__uni_bi	0.8047
MultinomialNB	{'alpha': 1}	corpus_estres__binario__uni_bi	0.8025
MultinomialNB	{'alpha': 1}	corpus_estres__frecuencia__uni_tri	0.8023
MultinomialNB	{'alpha': 0.1}	corpus_estres__binario__uni_tri	0.8
MultinomialNB	{'alpha': 1}	corpus_estres__frecuencia__uni_bi	0.7983
LogisticRegression	{'C': 10}	corpus_estres__binario__uni_bi	0.7982
MultinomialNB	{'alpha': 0.1}	corpus_estres__tfidf__uni_tri	0.7982
MultinomialNB	{'alpha': 1}	corpus_estres__binario__unigramas	0.7961
MultinomialNB	{'alpha': 0.1}	corpus_estres__tfidf__unigramas	0.7916
MultinomialNB	{'alpha': 0.1}	corpus_estres__frecuencia__unigramas	0.79
SVC	{'C': 10, 'kernel': 'rbf'}	corpus_estres__binario__unigramas	0.7888
SVC	{'C': 0.1, 'kernel': 'linear'}	corpus_estres__binario__uni_bi	0.7886
SVC	{'C': 0.1, 'kernel': 'linear'}	corpus_estres__frecuencia__uni_bi	0.784
LogisticRegression	{'C': 1}	corpus_estres__frecuencia__uni_bi	0.7839
SVC	{'C': 10, 'kernel': 'rbf'}	corpus_estres__tfidf__unigramas	0.7834
LogisticRegression	{'C': 10}	corpus_estres__tfidf__unigramas	0.7819
LogisticRegression	{'C': 10}	corpus_estres__tfidf__uni_tri	0.7818
SVC	{'C': 0.1, 'kernel': 'linear'}	corpus_estres__binario__uni_tri	0.7813
LogisticRegression	{'C': 1}	corpus_estres__frecuencia__uni_tri	0.7811
LogisticRegression	{'C': 1}	corpus_estres__binario__unigramas	0.7793
LogisticRegression	{'C': 1}	corpus_estres__binario__uni_tri	0.7783
SVC	{'C': 0.1, 'kernel': 'linear'}	corpus_estres__frecuencia__uni_tri	0.7771
SVC	{'C': 10, 'kernel': 'rbf'}	corpus_estres__frecuencia__unigramas	0.7759
LogisticRegression	{'C': 1}	corpus_estres__frecuencia__unigramas	0.7718
RandomForest	{'max_depth': 20, 'n_estimators': 100}	corpus_estres__frecuencia__unigramas	0.7509
MultinomialNB	{'alpha': 10}	corpus_estres__binario__bigramas	0.7459
MultinomialNB	{'alpha': 10}	corpus_estres__frecuencia__bigramas	0.7435
RandomForest	{'max_depth': 20, 'n_estimators': 100}	corpus_estres__binario__unigramas	0.7421
RandomForest	{'max_depth': 20, 'n_estimators': 100}	corpus_estres__tfidf__unigramas	0.7411
RandomForest	{'max_depth': 20, 'n_estimators': 200}	corpus_estres__binario__uni_bi	0.7383
SVC	{'C': 1, 'kernel': 'linear'}	corpus_estres__tfidf__bigramas	0.7301
SVC	{'C': 0.1, 'kernel': 'linear'}	corpus_estres__frecuencia__bigramas	0.7281
SVC	{'C': 0.1, 'kernel': 'linear'}	corpus_estres__binario__bigramas	0.7262

RandomForest	{'max_depth': None, 'n_estimators': 100}	corpus_estres__frecuencia__uni_bi	0.7258
MultinomialNB	{'alpha': 0.1}	corpus_estres__tfidf__bigramas	0.7133
RandomForest	{'max_depth': 20, 'n_estimators': 200}	corpus_estres__tfidf__uni_bi	0.7105
LogisticRegression	{'C': 10}	corpus_estres__tfidf__bigramas	0.71
LogisticRegression	{'C': 10}	corpus_estres__binario__bigramas	0.7094
RandomForest	{'max_depth': 20, 'n_estimators': 200}	corpus_estres__binario__uni_tri	0.7092
LogisticRegression	{'C': 1}	corpus_estres__frecuencia__bigramas	0.7069
RandomForest	{'max_depth': 20, 'n_estimators': 200}	corpus_estres__frecuencia__uni_tri	0.7036
RandomForest	{'max_depth': 20, 'n_estimators': 200}	corpus_estres__tfidf__uni_tri	0.69
MultinomialNB	{'alpha': 0.1}	corpus_estres__tfidf__trigramas	0.562
MultinomialNB	{'alpha': 1}	corpus_estres__binario__trigramas	0.5471
MultinomialNB	{'alpha': 1}	corpus_estres__frecuencia__trigramas	0.5471
SVC	{'C': 10, 'kernel': 'rbf'}	corpus_estres__binario__trigramas	0.4955
SVC	{'C': 10, 'kernel': 'rbf'}	corpus_estres__frecuencia__trigramas	0.4955
RandomForest	{'max_depth': 20, 'n_estimators': 100}	corpus_estres__tfidf__bigramas	0.4847
RandomForest	{'max_depth': 20, 'n_estimators': 200}	corpus_estres__binario__bigramas	0.4846
RandomForest	{'max_depth': 20, 'n_estimators': 200}	corpus_estres__frecuencia__bigramas	0.4751
SVC	{'C': 10, 'kernel': 'linear'}	corpus_estres__tfidf__trigramas	0.4656
LogisticRegression	{'C': 10}	corpus_estres__binario__trigramas	0.4105
LogisticRegression	{'C': 10}	corpus_estres__frecuencia__trigramas	0.4105
LogisticRegression	{'C': 10}	corpus_estres__tfidf__trigramas	0.4059
RandomForest	{'max_depth': 20, 'n_estimators': 200}	corpus_estres__binario__trigramas	0.3581
RandomForest	{'max_depth': None, 'n_estimators': 100}	corpus_estres__tfidf__trigramas	0.3572
RandomForest	{'max_depth': None, 'n_estimators': 100}	corpus_estres__frecuencia__trigramas	0.3533

Matrices de confusión los mejores 3 modelos






Confusión: corpus_estres__tfidf__uni_bi.pkl | MultinomialNB
51
14
8
49
no
si
Valor real
no
si
Predicción
50
45
40
35
30
25
20
15
10