Logotipo, nombre de la empresa

Descripción generada automáticamente Logotipo, nombre de la empresa

Descripción generada automáticamente

PIA

**Modelos de Clasificación**

**Materia: MAA(Aprendizaje Automático)**

**Profesor: José Anastacio Hernández Saldaña**

**Grupo: 03**

**Alumno: José de Jesús Almanza Trejo**

Contenido

[Introducción 3](#_Toc173019115)

[Análisis descriptivo 4](#_Toc173019116)

[Clasificación 5](#_Toc173019117)

[Resultados 6](#_Toc173019118)

[Bibliografía 8](#_Toc173019119)

# Introducción

A continuación se presenta una serie de modelos aplicados a un set de información referente a natural language processing model (se deduce por el nombre de las features), donde la característica referente a engagement es la variable objetivo (una variable binaria).

Se realizará la búsqueda de un mejor modelo aplicando modelos de clasificación (binaria) al conjunto de datos tal que el valor del ROC AUC (Area Under de ROC Curve) se maximice, utilizando una herramienta (optuna) de búsqueda óptima de parámetros para los diferentes algoritmos.

Una vez encontrado el mejor modelo se presentarán su algoritmo y parámetros.

# Análisis descriptivo

Tenemos las siguientes características de las variables en el dataset:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Conteo | Promedio | Mínimo | Percentil 25% | Percentil 50% | Percentil 75% | Máximo |
| title\_word\_count | 9,239 | 7.70 | 1.00 | 5.00 | 7.00 | 10.00 | 33.00 |
| document\_entropy | 9,239 | 7.79 | 0.00 | 7.59 | 7.88 | 8.16 | 9.28 |
| freshness | 9,239 | 14,808.59 | 10,830.00 | 14,070.00 | 14,750.00 | 15,600.00 | 17,430.00 |
| easiness | 9,239 | 84.76 | 28.21 | 80.41 | 84.48 | 88.39 | 122.03 |
| fraction\_stopword\_presence | 9,239 | 0.61 | 0.00 | 0.59 | 0.61 | 0.63 | 1.00 |
| normalization\_rate | 9,239 | 0.02 | 0.00 | 0.01 | 0.02 | 0.03 | 0.10 |
| speaker\_speed | 9,239 | 2.41 | 0.00 | 1.98 | 2.27 | 2.54 | 50.85 |
| silent\_period\_rate | 9,239 | 0.15 | 0.00 | 0.00 | 0.10 | 0.25 | 1.17 |

Se muestra el histograma de las variables:

Gráfico, Histograma

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico, Histograma

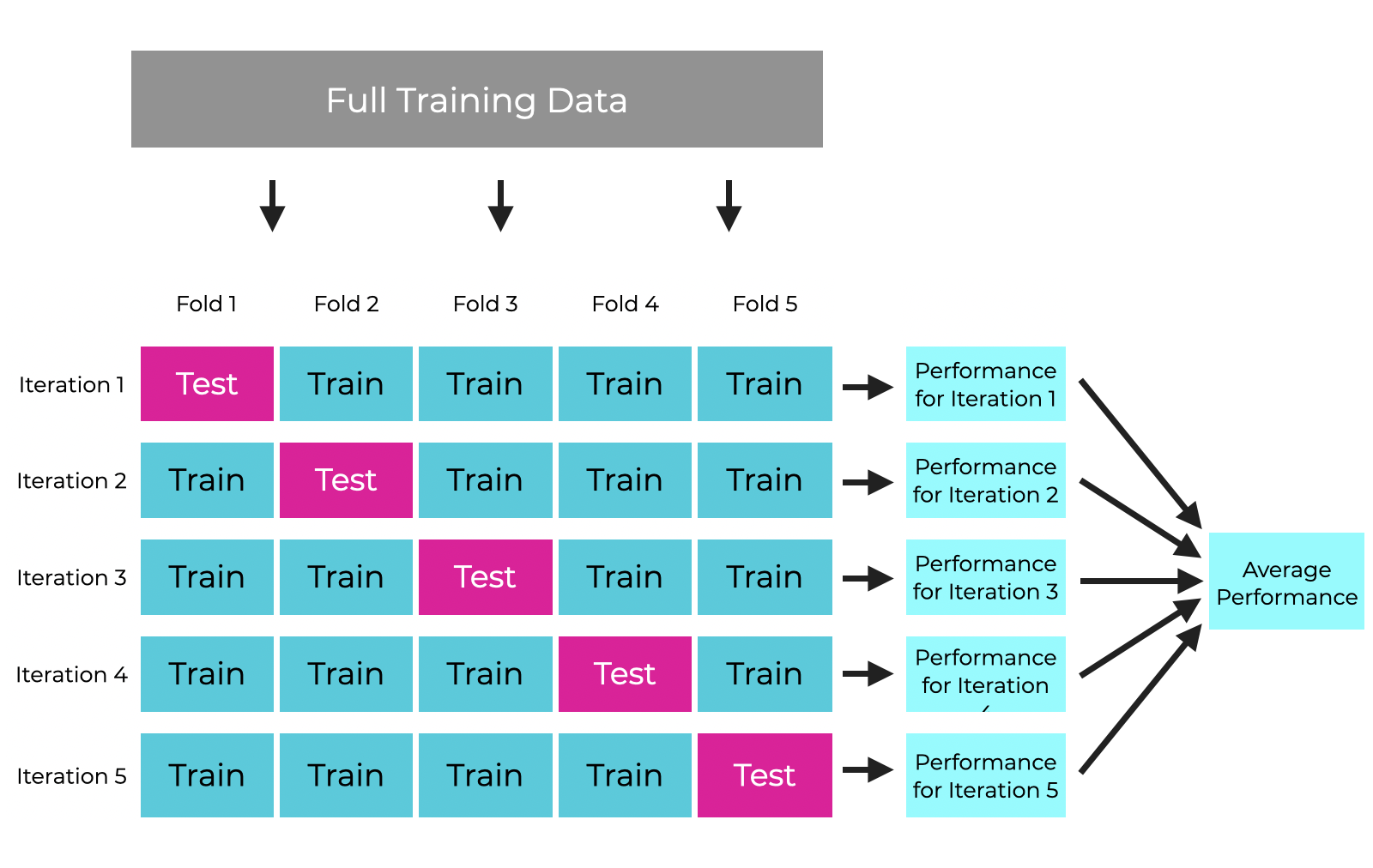
Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamenteGráfico, Histograma

Descripción generada automáticamente

# Clasificación

Se realizan varios ejercicios de regresión con diversos algoritmos vistos en clase así como otros tipos de algoritmos de clasificación que tiene la librería sklearn (Sklearn, 2024), el modelado consistirá de la aplicación de un algoritmo de validación cruzada a partir de 10 particiones estratificadas de la base de entrenamiento en las cuales se correrá el algoritmo seleccionado con sus parámetros tomando cada una de las particiones como set de pruebas y el resto de las particiones como set de entrenamiento, se calcula el score ROC AUC para cada iteración y al final se promedia el score para mostrar el desempeño promedio del modelo.



## Resultados

Se corre un algoritmo de búsqueda optimizada de parámetros por medio de la librería optuna (Optuna, 2018) en donde se buscan los mejores parámetros que maximicen nuestro score seleccionado (ROC AUC) obteniendo los siguientes resultados:

|  |  |  |
| --- | --- | --- |
| Algoritmo | Parámetros | ROC AUC promedio |
| Logistic Regression | {'penalty': 'l2', 'solver': 'liblinear', 'C': 0.001, 'l1\_ratio': None} | 0.8476 |
| Decision Tree Classifier | {'max\_depth': 44, 'criterion': 'log\_loss', 'min\_samples\_split': 0.024854806445273624, 'min\_samples\_leaf': 0.026777469908052437} | 0.8642 |
| Random Forest Classifier | {'n\_estimators': 122, 'max\_depth': 179, 'criterion': 'entropy', 'min\_samples\_split': 0.010022356623229676, 'min\_samples\_leaf': 0.01014036579990787, 'max\_leaf\_nodes': 42} | 0.8791 |
| **Gradient Boosting Classifier** | **{'learning\_rate': 0.21953581594646793, 'n\_estimators': 57, 'max\_depth': 26, 'subsample': 0.978232013339438, 'min\_samples\_split': 0.2641324973039342, 'min\_samples\_leaf': 0.04570519683574234, 'max\_leaf\_nodes': 20}** | **0.8978** |

El mejor modelo seleccionado es un Gradient Boosting Classifier con random\_seed=123 y arroja un score promedio de ROC AUC de 0.8978.

Se muestran los gráficos de ROC AUC de los algoritmos:

|  |  |
| --- | --- |
| Gradient Boosting | Random Forest |
| Diagrama  Descripción generada automáticamente | Diagrama  Descripción generada automáticamente |
| Regresión Logística | Decision Tree |
|  |  |

# Bibliografía

Optuna. (2018). Obtenido de https://optuna.readthedocs.io/en/stable/tutorial/index.html

Sklearn. (2024). Obtenido de https://scikit-learn.org/stable/supervised\_learning.html