Resultados Etapa 1

Relacionamiento automático de opiniones

Juan Felipe Lancheros
Obed Cabanzo
Santiago Tapias

Agenda

1. Preparación de datos
2. Modelos
3. Metricas y comparación
4. Modelo seleccionado
5. Resultados

Preparación de datos

Perfilamiento

Se trabaja con un conjunto de datos que comprende 4049 registros asociados a la clasificación de información brindada por ciudadanos según un ODS: 3, 4 y 5.

No hay problemas asociados a completitud.

Calidad y tratamiento

Se aplicó eliminación de ruido a todos los registros mediante el preprocesamiento de los datos (Remoción de NON-ASCII caracteres, eliminación de puntuación, stop words, lematización, etc).

Finalmente, se aplicó TF_IDF para resaltar numéricamente aquellos términos que son más representativos del contenido de las opiniones.

Modelos

Gaussian Naive Bayes

Fue elegido para esta tarea debido a su capacidad para manejar eficientemente problemas de clasificación multiclase con datos continuos, como los vectores resultantes del TF-IDF.

Árbol de decisión

Aparte de su flexibilidad en la preparación de los datos para modelos de clasificación, su jerarquía brinda facilidad de entendimiento e interpretación, lo que podría ser útil para los stakeholders.

Máquinas de Vectores de Soporte

Es de alta precisión y capacidad de manejar clases desbalanceadas, garantizando un rendimiento consistente. Esto permite a la empresa clasificar de manera eficiente las opiniones según los ODS y apoyando la toma de decisiones estratégicas.

Ventajas

Gaussian Naive Bayes

El modelo Naive Bayes, encontrado a partir de un proceso de cross-validation para determinar los mejores parámetros para TD-IDF.

Árbol de decisión

Gracias a los hiperpárametros es posible hallar diferentes configuraciones que apunten a los mejores indicadores estadísticos

Máquinas de Vectores de Soporte

Tiene capacidad para manejar datos de alta dimensionalidad y generar un margen óptimo de separación entre clases

SVM

Eficiencia

Capacidad

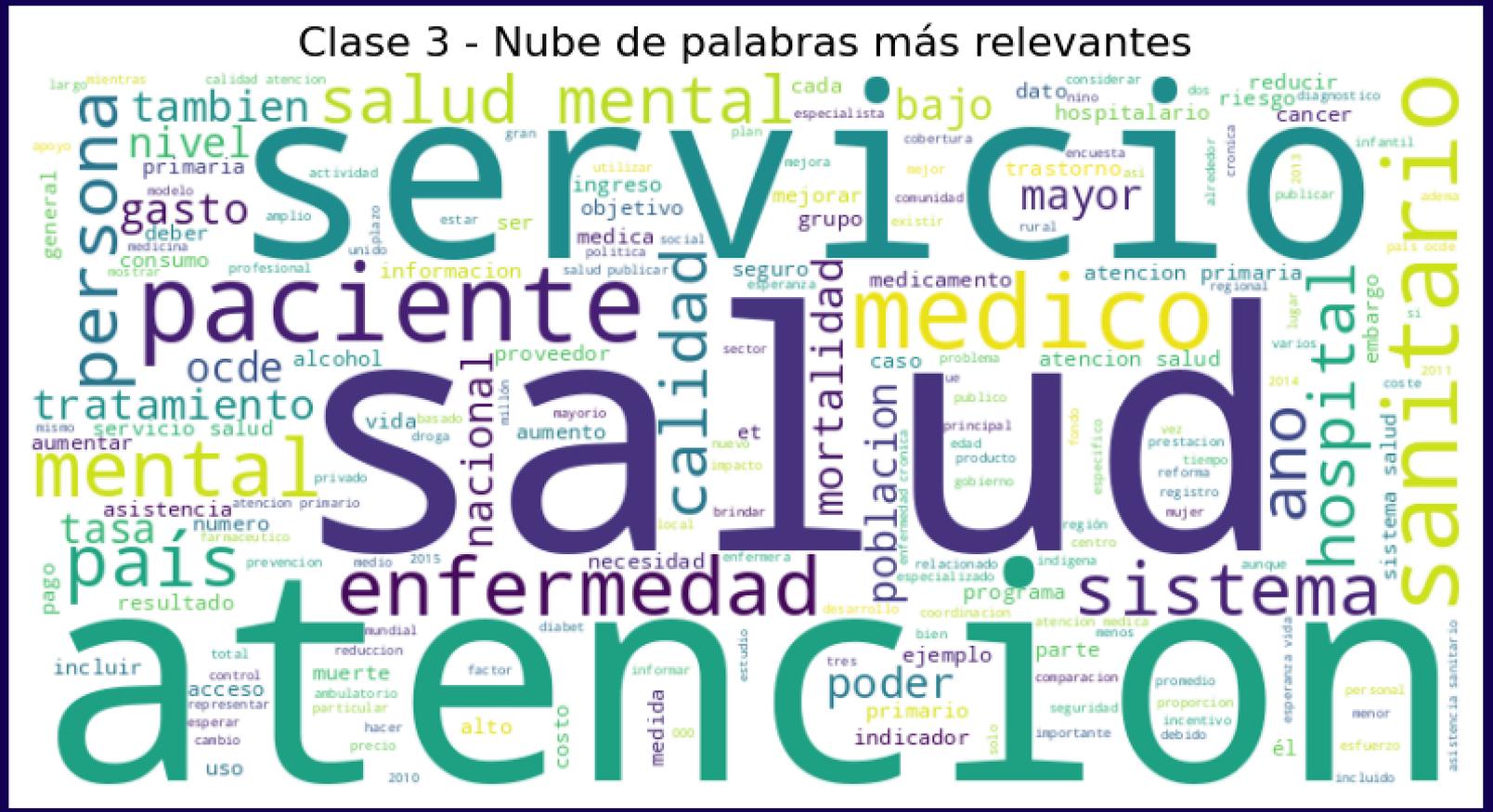
Generalización

Optimización

Selección del modelo + Métricas

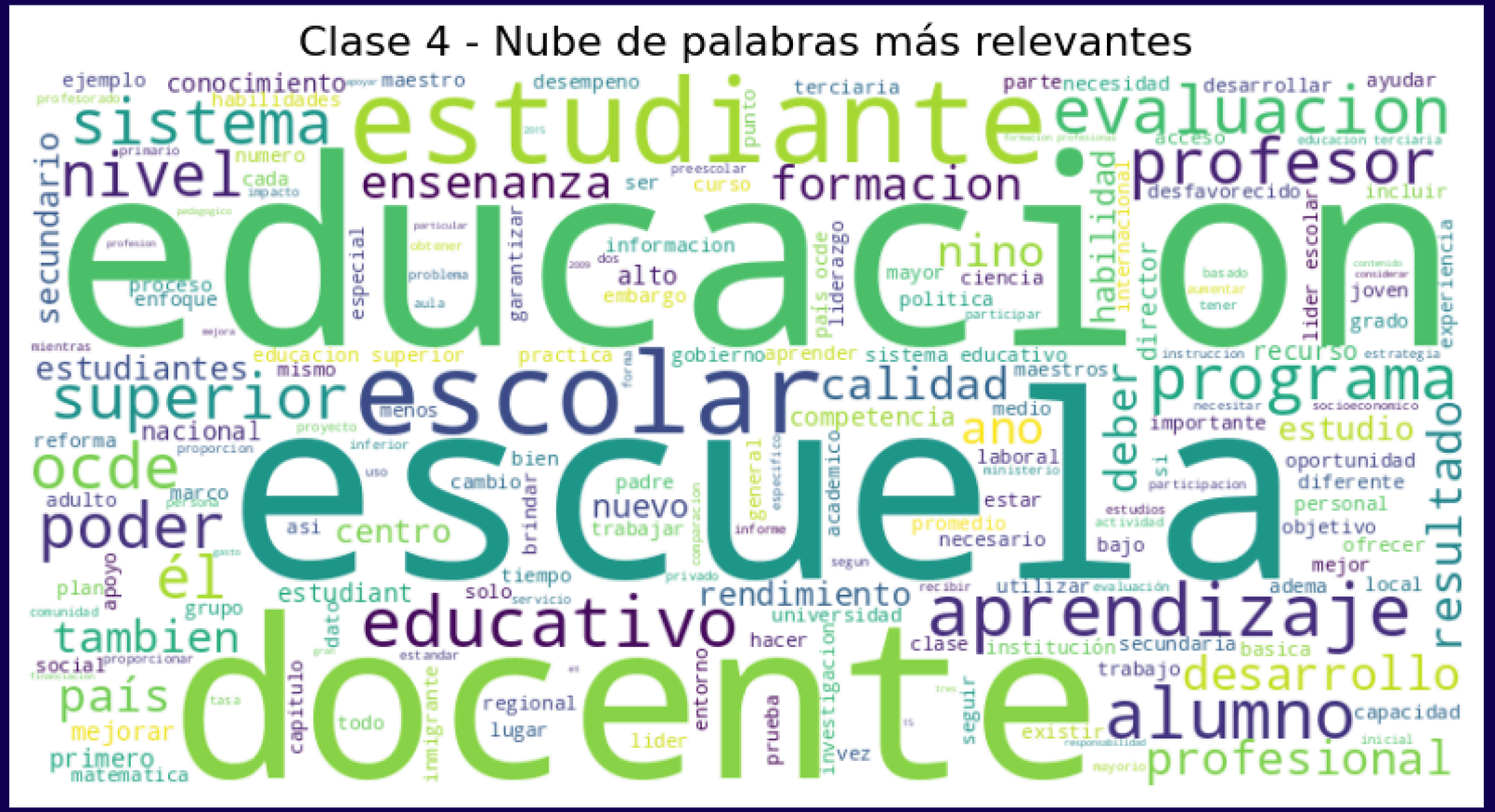
Las siguientes métricas son indicadores estadísticos que evalúan el desempeño de los modelos sobre el conjunto de datos. Con sus valores, determinaremos el mejor para el negocio.

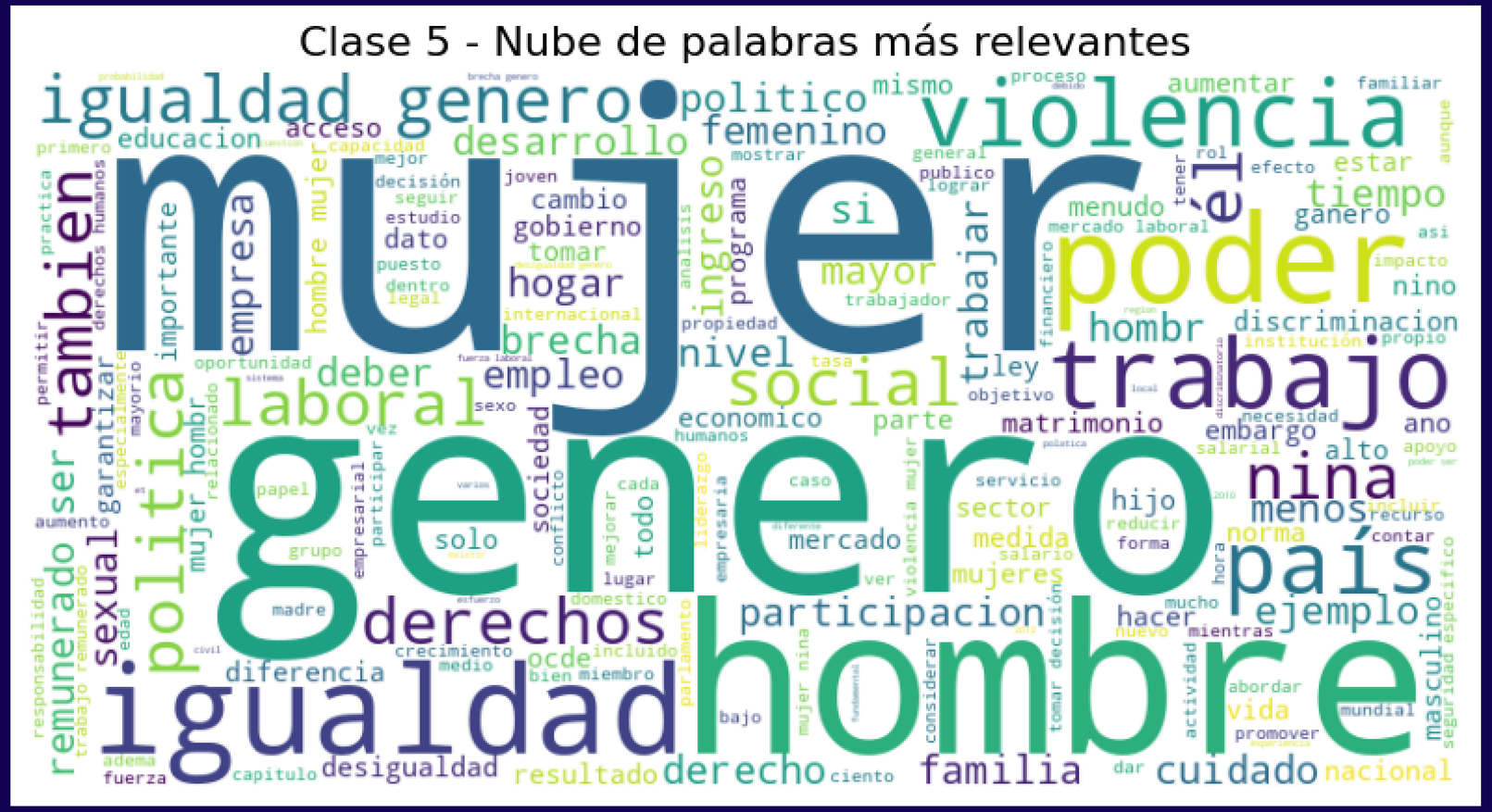
	Precision	Recall	F1-score	Support
GNB	0.95	0.94	0.94	810
AD	0.93	0.93	0.93	810
SVM	0.98	0.98	0.98	1620



Palabras Clave
Clase 3

Palabras Clave
Clase 4





Palabras Clave Clase 5

Universidad de Los Andes

Muchas gracias