

ANÁLISIS INTELIGENTE DE DATOS:

TAREA 3

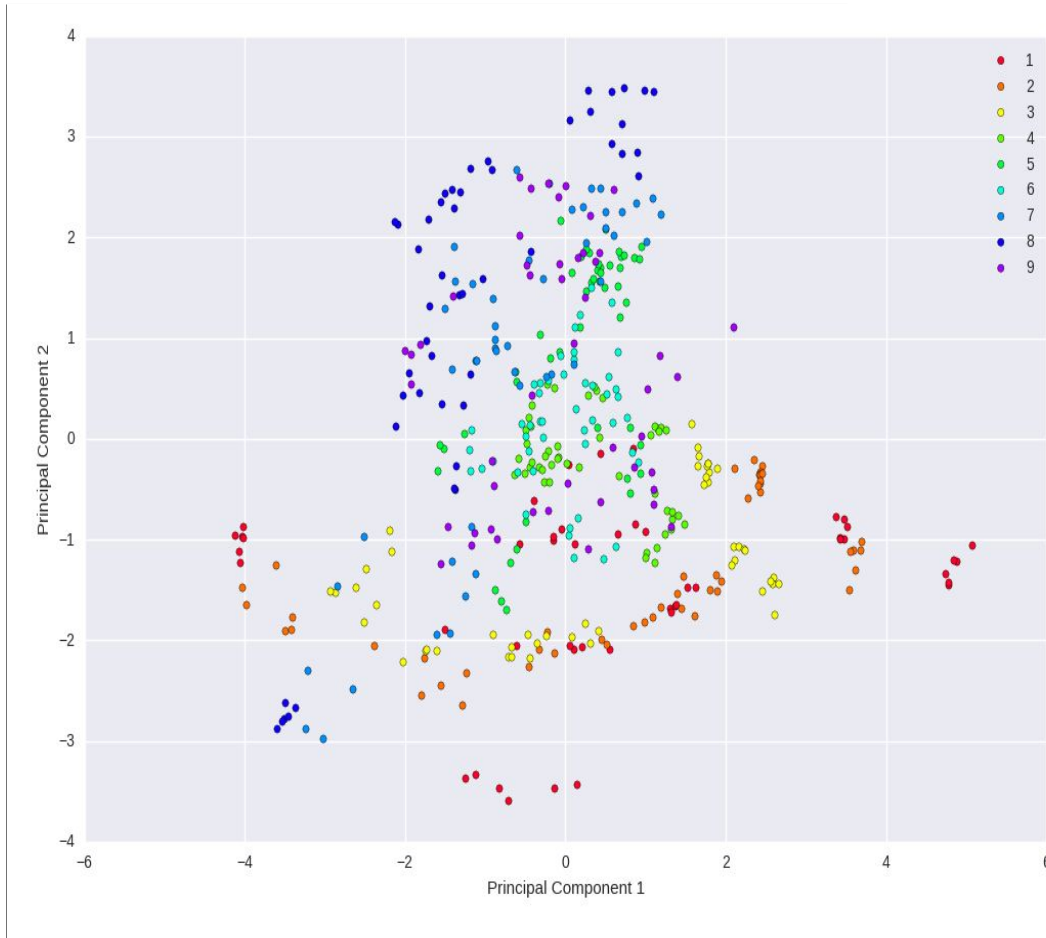
Diego Salazar
Anibal Pérez

PARTE 1: INTRODUCCIÓN

Se utilizarán distintos métodos de reducción de dimensionalidad, para analizar su efecto en la **precisión y error** de predicción sobre un dataset de sonidos fonéticos con 10 características, para su predicción sobre 11 vocales del alfabeto británico:

- Train-set : 528 registros, 48 veces c/u.
- Test-set : 462 registros, 42 veces c/u.
- Características: 10 variables numéricas discretas.

PRINCIPAL COMPONENT ANALYSIS (PCA)



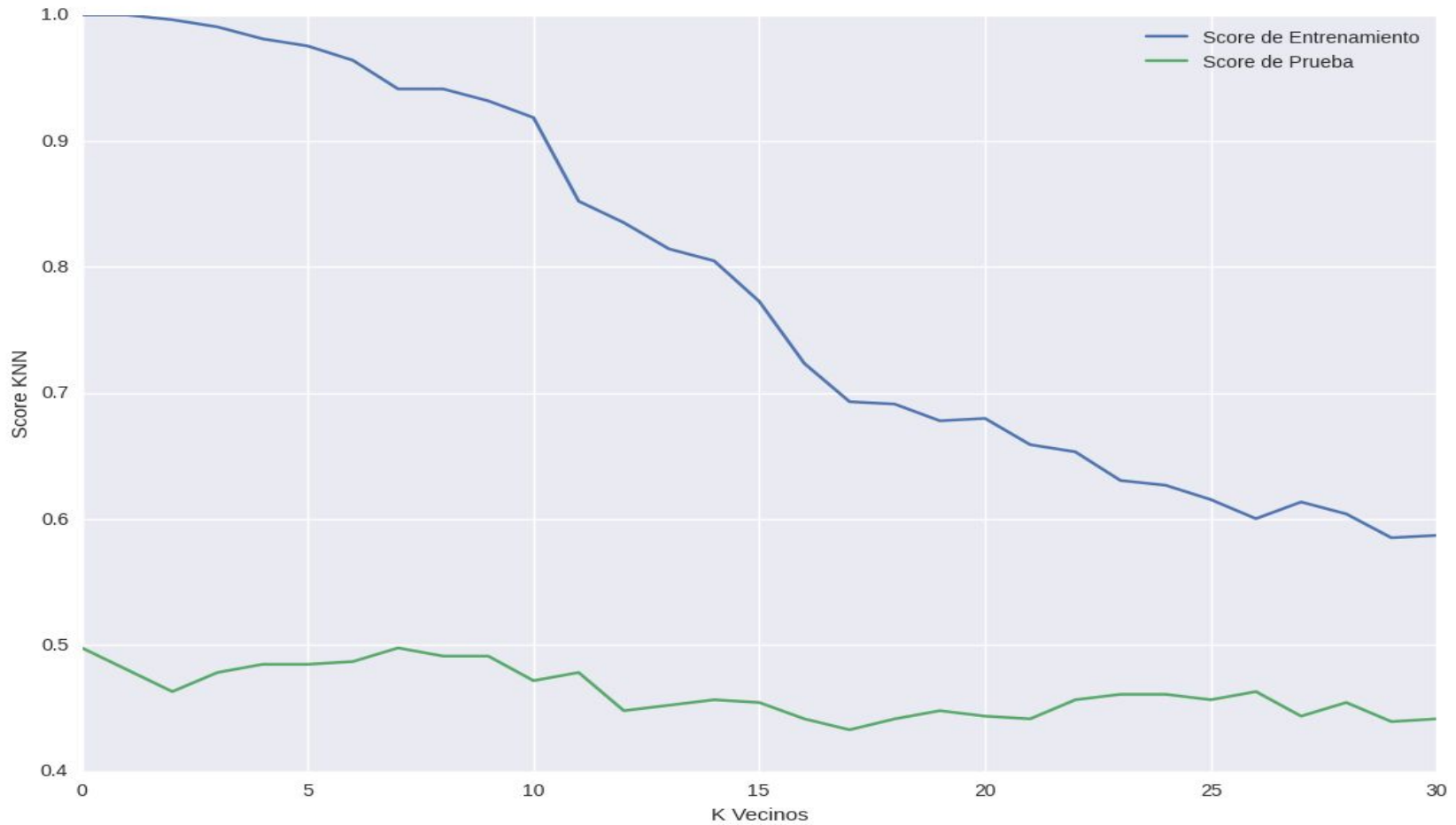
- Paleta arcoiris.
- Se pueden distinguir datos de distintas clases.
- Se observa mucha dispersión, colores poco agrupados.
- PCA codifica sin mirar respuesta.

LINEAR DISCRIMINANT ANALYSIS (LDA)



- Se aprecia una mejor agrupación de colores.
- Mejor clasif. pues se basa en similitudes.
- Regresión lineal con respecto a la respuesta, se esperaba mejor comportamiento.

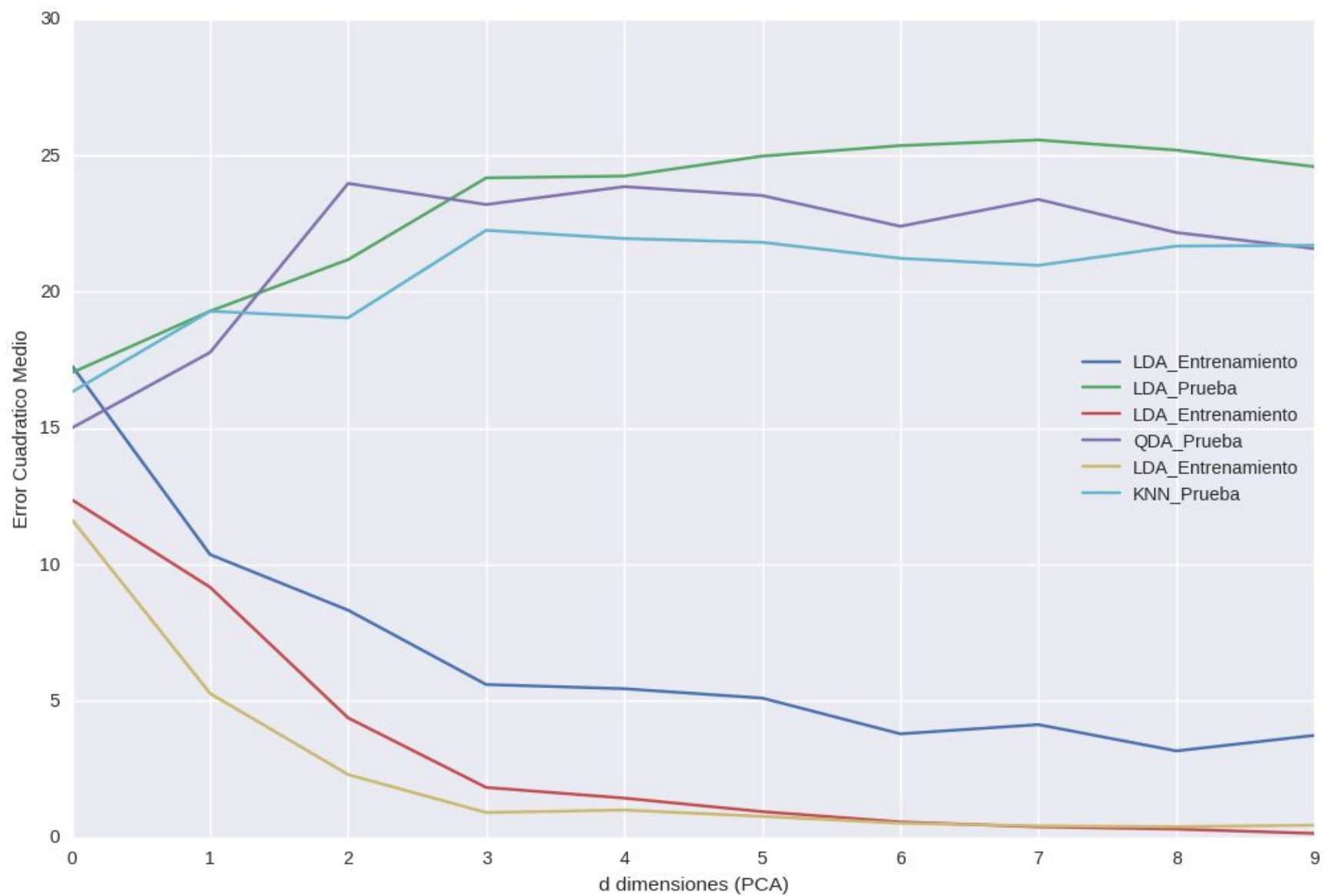
DESEMPEÑO DE KNN SEGÚN K-VECINOS

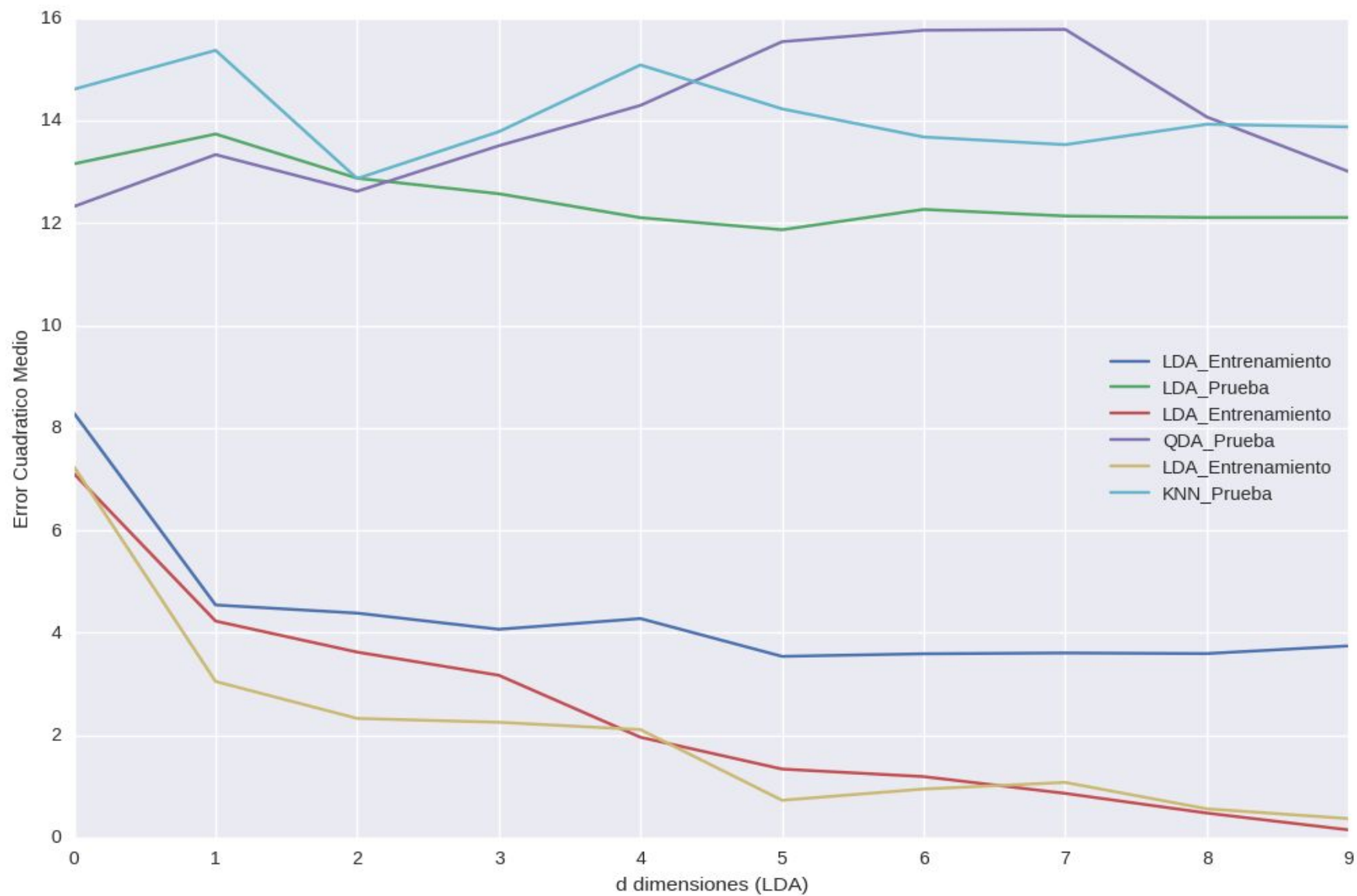


COMPARACIÓN DE DESEMPEÑO

| Técnica Utilizada\Tipo-Set | Set Entrenamiento | Set de Pruebas |
|----------------------------|-------------------|----------------|
| LDA | 0.6837 | 0.4524 |
| QDA | 0.9886 | 0.4156 |
| KNN (n=10) | 0.9318 | 0.4918 |

- QDA se comporta mejor que LDA y KNN para el set de entrenamiento.
- KNN se comporta mejor que LDA y QDA para el set de pruebas.





PARTE 2: INTRODUCCIÓN

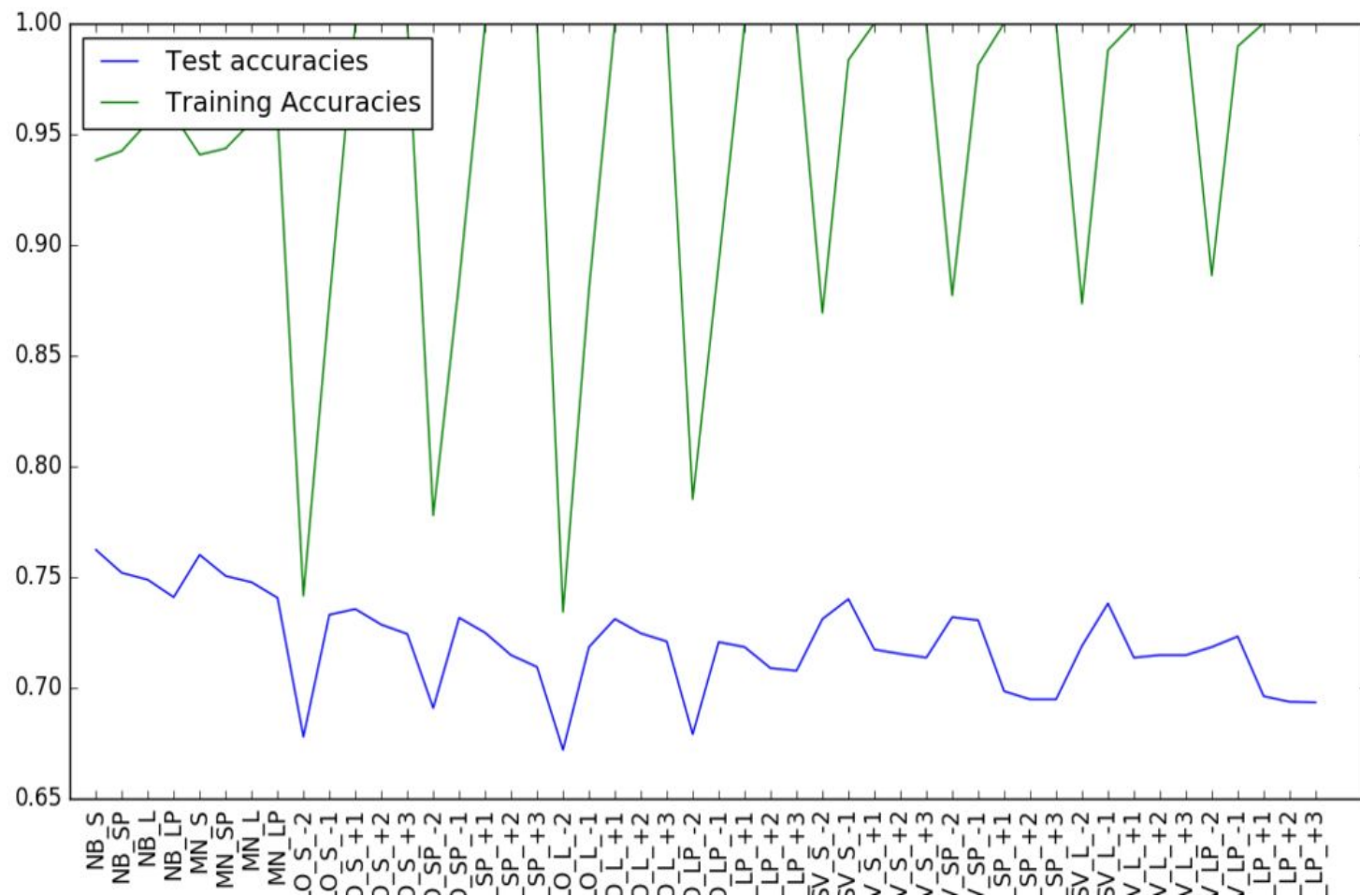
Se propone la creación de modelos predictivos para realizar un análisis de sentimiento sobre data textual -particularmente, críticas de películas. Para esto se utilizarán clasificadores binarios y se comparará sus poderes predictivos para esta situación.

CARACTERÍSTICAS DE LA DATA

Los datos serán procesados de 4 maneras:

- Stemming sin preprocesamiento
- Stemming con preprocesamiento
- Lematización sin preprocesamiento
- Lematización con preprocesamiento

RESULTADOS



GRACIAS POR SU ATENCIÓN!