Análisis Inteligente de Datos - Tarea 3

Iván González - Felipe Vásquez

Reducción de dimensionalidad para clasificación

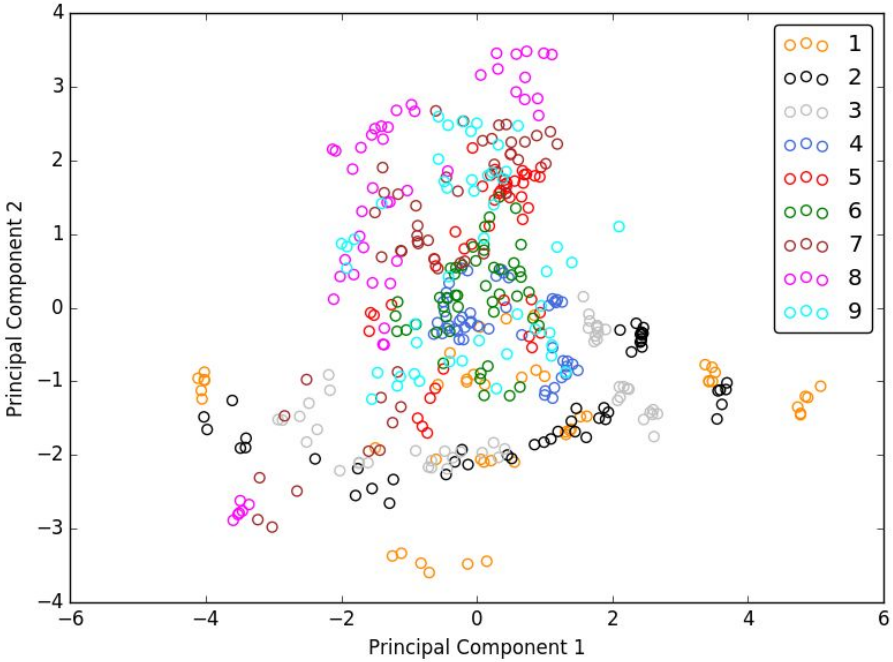
Conjunto de Datos

- Reconocimiento de vocales
- 11 clases, 10 predictores

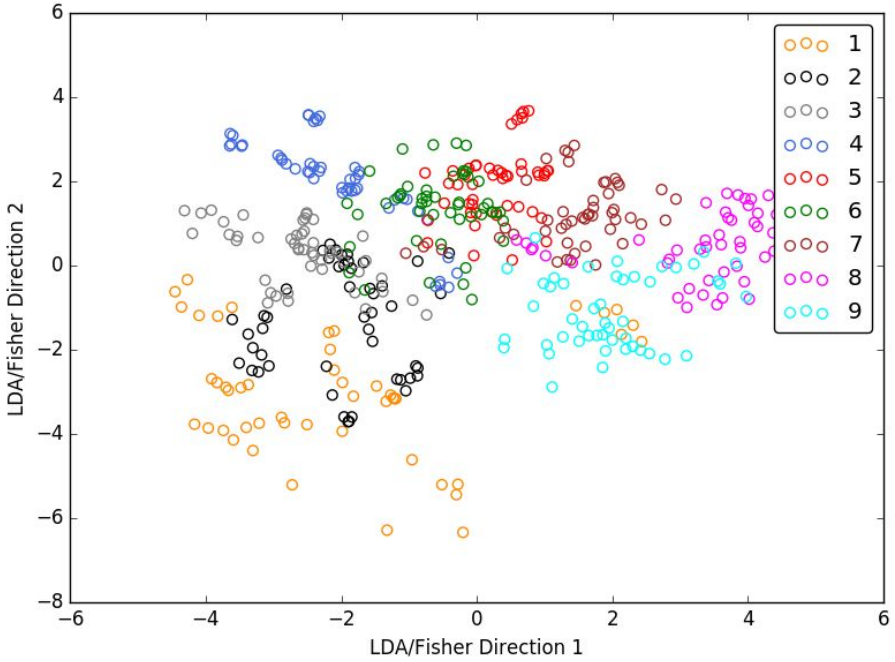
Vowel	Word	Vowel	Word	Vowel	Word	Vowel	Word
i:	heed	O	hod	I	hid	C:	hoard
E	head	U	hood	A	had	u:	who'd
a:	hard	3:	heard	Y	hud		

- Los dataset de entrenamiento y de pruebas están compuestos de 518 y 462 registros respectivamente.

Representación de 2 dimensiones



PCA

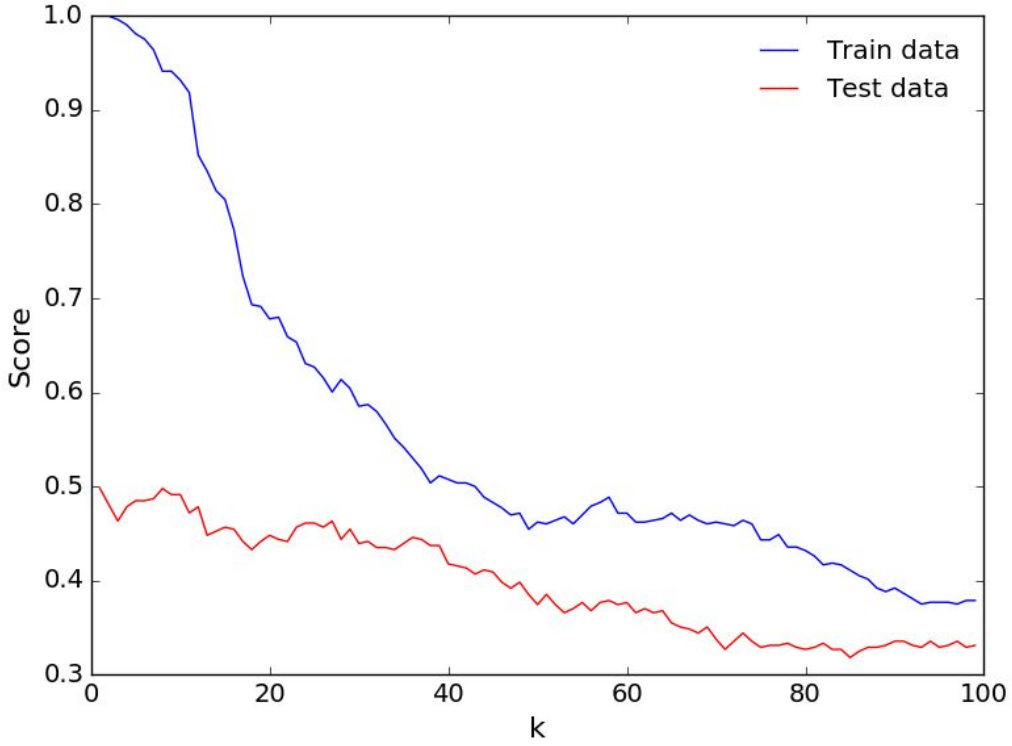


LDA

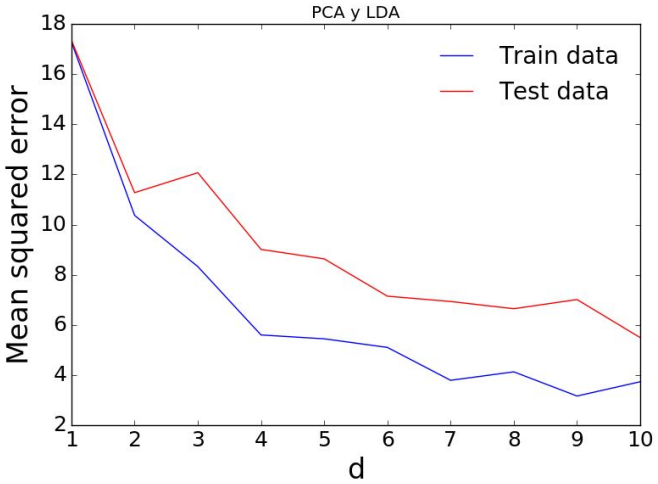
Comparación LDA, QDA y KNN

Modelo	**Score datos de entrenamiento**	**Score datos de pruebas**
LDA	0.684	0.452
QDA	0.969	0.416
KNN *(k=10)*	0.932	0.491

KNN: Score en función del número de vecinos

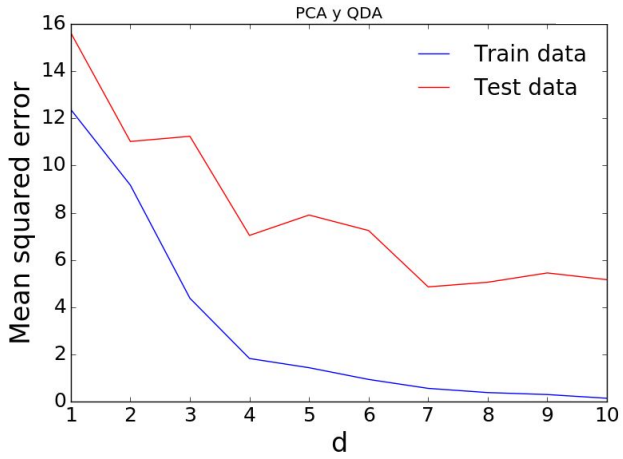


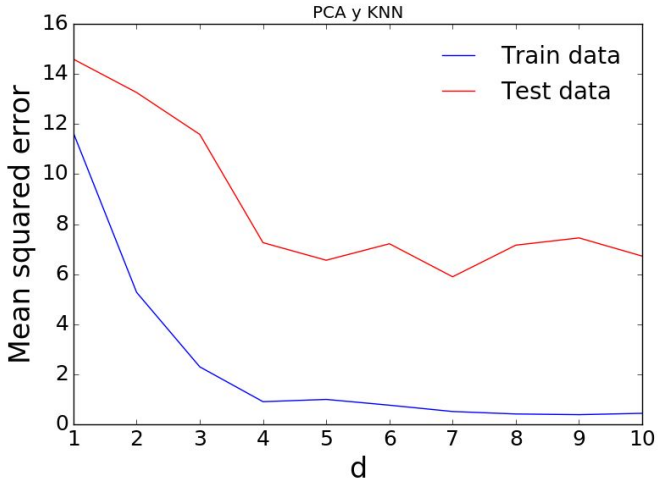
Reducción de dimensionalidad con PCA



LDA

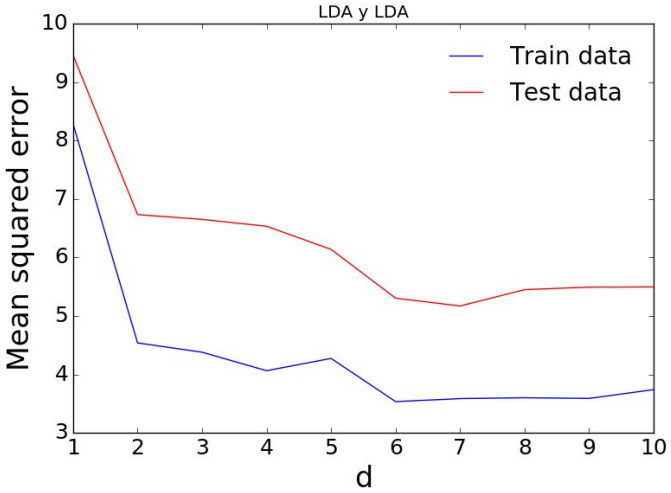
QDA



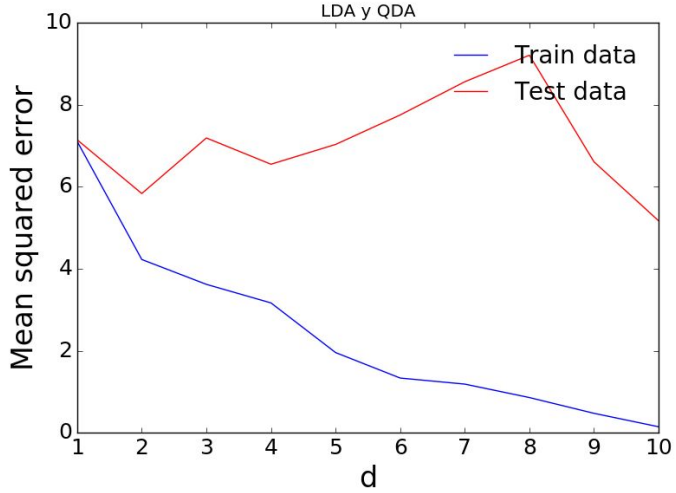


KNN *(k=10)*

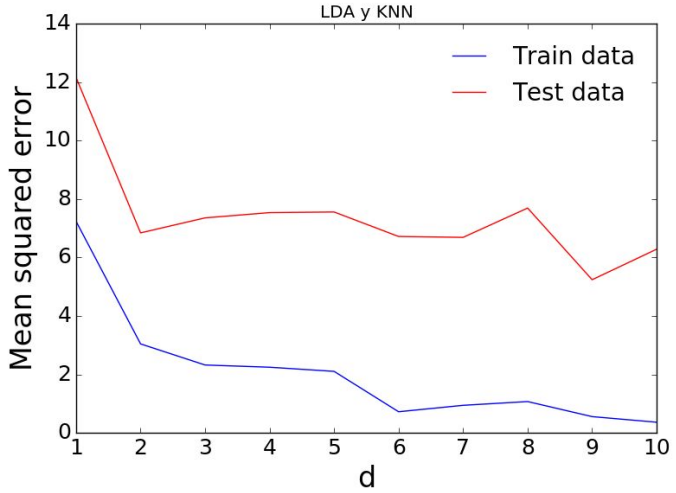
Reducción de dimensionalidad con LDA



LDA



QDA



KNN *(k=10)*

Comparación de los MSE mínimos test dataset

Modelo de clasificación	Min MSE test	Dimensiones
LDA	5,498	10
QDA	4,864	7
KNN *(k=10)*	5,900	7

Reducción de dimensionalidad con PCA

Modelo de clasificación	Min MSE test	Dimensiones
LDA	5,171	7
QDA	5,169	10
KNN *(k=10)*	5,240	9

Reducción de dimensionalidad con LDA

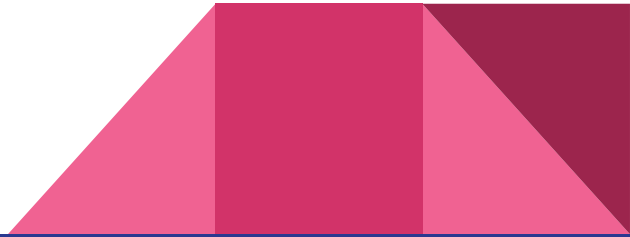
Análisis de opiniones sobre películas

Conjunto de Datos

- Tanto el conjunto de Entrenamiento como el de Pruebas tienen 3554 registros para cada clase.
- La variable de clase es Sentiment, que refleja la opinión positiva (1) o negativa (0) de un comentario de una persona.
- El texto (opinión) se representa como un vector de características para ser utilizado como predictor.
  Para esto es necesario construir un vocabulario a partir del dataset y contar cuántas veces aparece las palabras en el vocabulario.

Stemming

- Stemming es un proceso heurístico que corta la derivación de las palabras en la raíz.
- Ejemplo de Stemming:
  - I love eating the cake: “love eat cake”.
  - I loved eating the cake: “love eat cake”.
- Ejemplo sin Stemming:
  - I love eating the cake: “love eating cake”.
  - I loved eating the cake: “loved eating cake”.



Lematización

- Lematización es similar a stemming, en el sentido que reducen las formas inflecciones de las palabras a una base común o raíz.
- Lematización además, realiza un chequeo de la forma reducida en el corpus de WordNet.
- Ejemplo de Lematización:
    - I love eating cake: “love eating cake”. -> “eating” no se reduce a “eat”.

Palabras más frecuentes

Palabras más frecuentes en el conjunto de Entrenamiento y Prueba utilizando lematización.

Entrenamiento		Prueba	
Frecuencia	**Palabra**	**Frecuencia**	**Palabra**
556	film	558	film
481	movie	540	movie
246	one	250	one
245	like	238	ha
224	ha	230	like
183	make	197	story
176	story	175	character
163	character	165	time
145	comedy	161	make
143	time	134	comedy

Métricas

Las métricas que calcula la función *classification_report* son la siguientes: Precisión, Recall y F1-score para cada clase

- Precisión: Cantidad de resultados positivos correctos divididos por la cantidad total de resultados positivos
- Recall: Cantidad de resultados positivos correctos dividido por el número de resultados positivos que se debería obtener.
- F1-score: Es una medida de la exactitud de una prueba, que se calcula como sigue:

$$F_1 = 2 \cdot \frac{Precisión \cdot Recall}{Precisión + Recall}$$

Clasificador Bayesiano Ingenuo Binario y Multinomial

Métricas	Lematización	L-sin stopwords	Stemming
Training Accuracy	0.959	0.955	0.943
Test Accuracy	0.739	0.749	0.748
Precisión	0.74	0.75	0.75
Recal	0.74	0.75	0.75
F1-score	0.74	0.75	0.75

Clasificador Bayesiano Ingenuo Binario.

Métricas	Lematización	L-sin stopwords	Stemming
Training Accuracy	0.959	0.955	0.942
Test Accuracy	0.741	0.748	0.750
Precisión	0.74	0.75	0.75
Recal	0.74	0.75	0.75
F1-score	0.74	0.75	0.75

Clasificador Bayesiano Ingenuo Multinomial.

Regresión Logística Regularizado y SVM Lineal

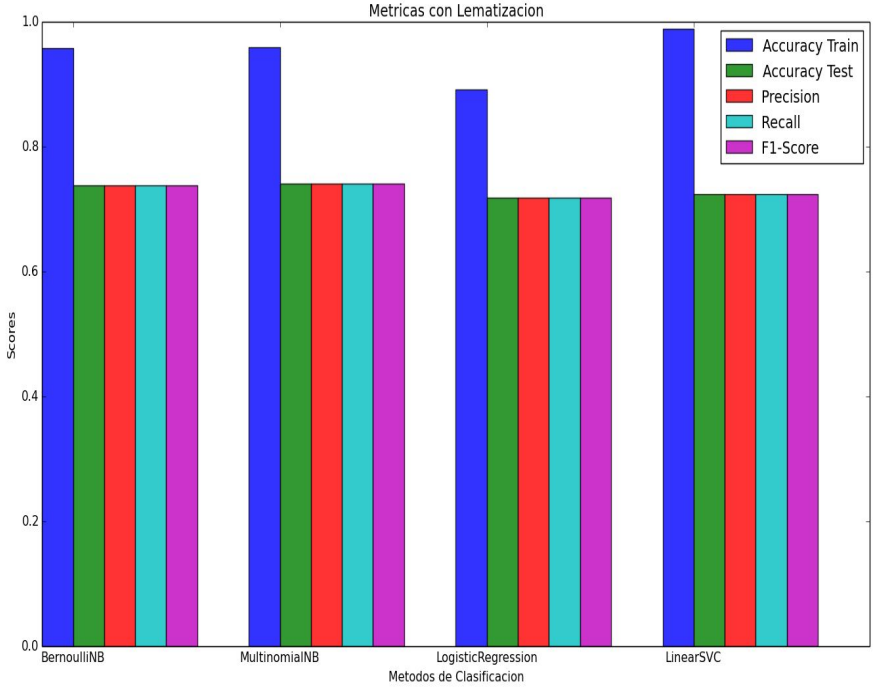
Métricas	Lematización	L-sin stopwords	Stemming
Training Accuracy	0.892	0.879	0.880
Test Accuracy	0.719	0.719	0.731
Precisión	0.72	0.72	0.73
Recal	0.72	0.72	0.73
F1-score	0.72	0.72	0.73

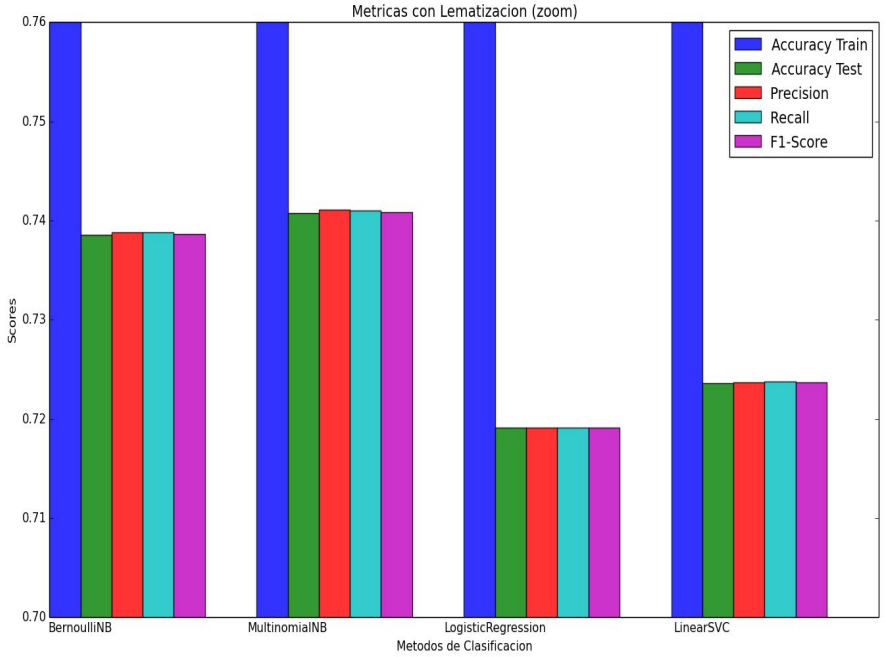
Regresión Logística con Regularización de 0.1.

Métricas	Lematización	L-sin stopwords	Stemming
Training Accuracy	0.989	0.988	0.982
Test Accuracy	0.724	0.738	0.731
Precisión	0.72	0.74	0.73
Recal	0.72	0.74	0.73
F1-score	0.72	0.74	0.73

SVM Lineal con Regularización de 0.1.

Comparación de métodos, con Lematización





Comparación de métodos, con Stemming