Máquinas de Aprendizaje
Tarea 2

Felipe Araya Barrera
Sebastián Vergara Miranda

Reducción de dimensionalidad para clasificación

Descripción de dataframes

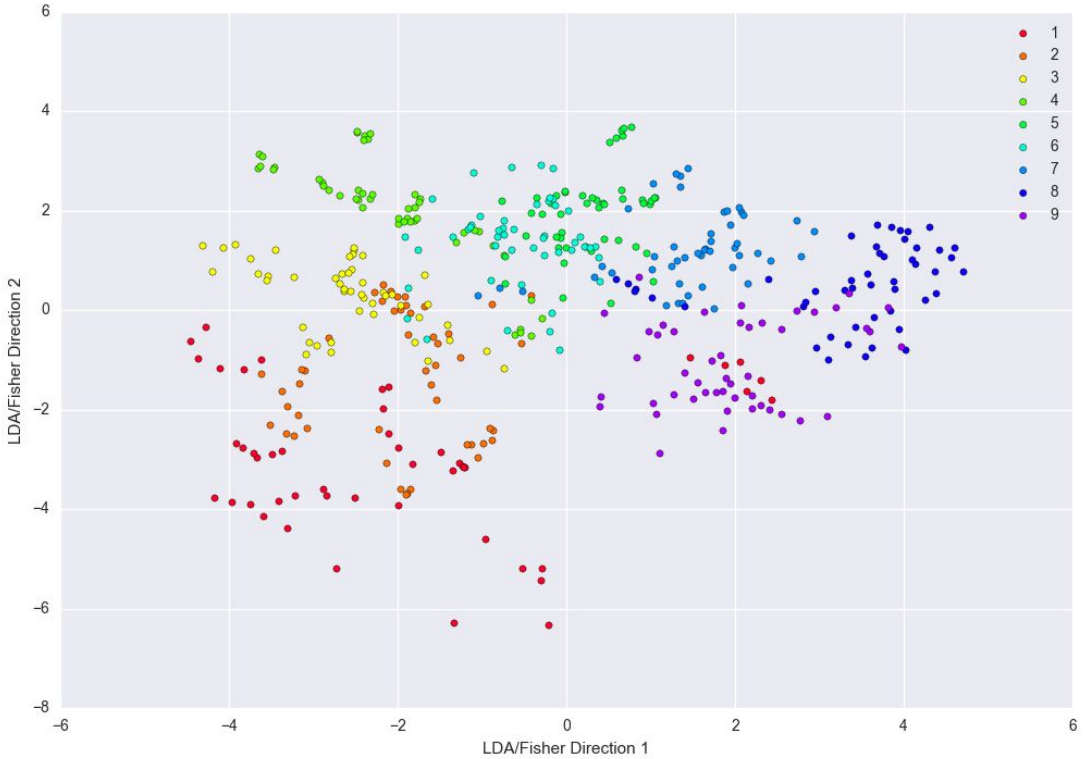
- **Datos de entrenamiento**: 528 registros.
- **Datos de prueba**: 462 registros.
- **Promedio de palabras por ítem en cada clase (conjunto de entrenamiento)**: 4.8 palabras.
- **Promedio de palabras por ítem en cada clase (conjunto de prueba)**: 4.2 palabras.

PCA (paleta *PuBuGn*)

PCA (paleta *gist_rainbow*)

LDA (paleta *PuBuGn*)

LDA (paleta *gist_rainbow*)



Construcción de clasificador

La construcción del clasificador se realizó sólo en base a la probabilidad de ocurrencia de cada clase:

$$P(C_i) = \frac{C_i}{Numero\_Datos\_DF}$$

En este caso todos las clases poseían la misma probabilidad, por ende para realizar la clasificación se optó por generar un número aleatorio para saber de que clase sería dato.

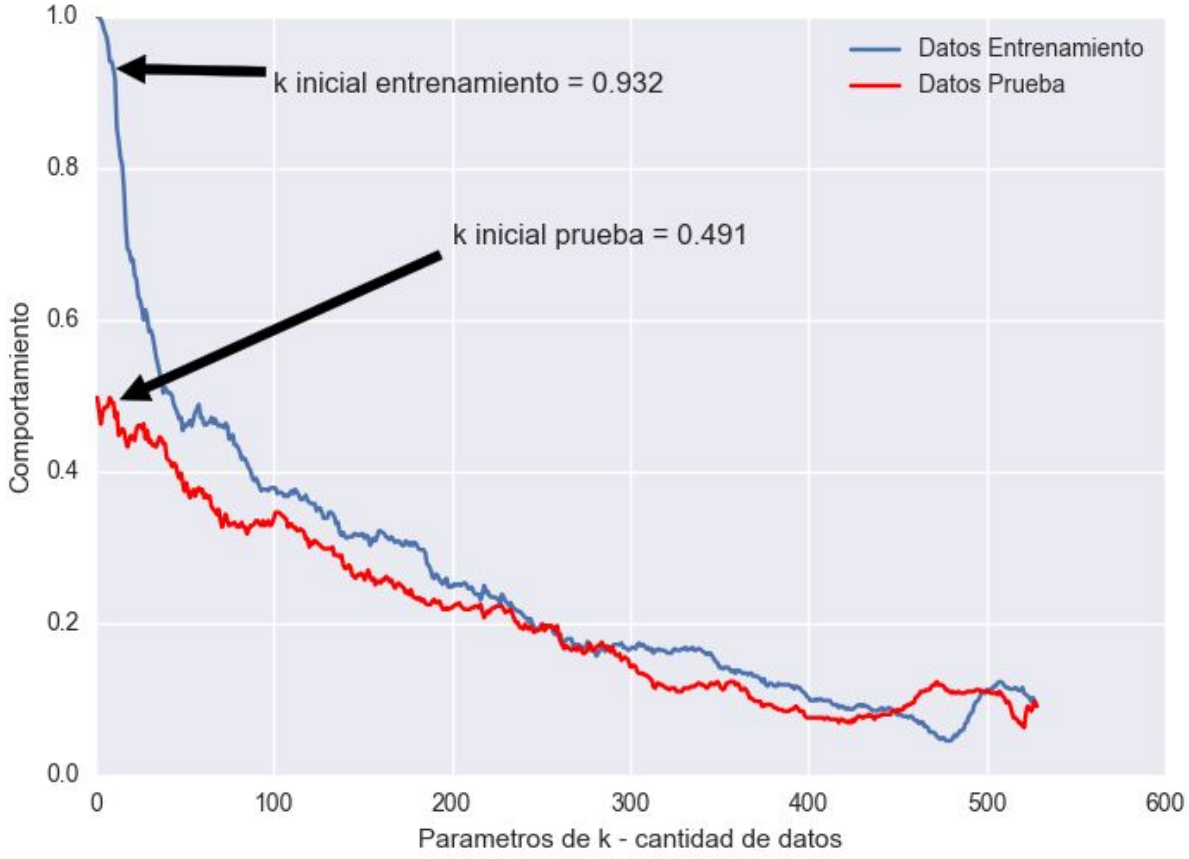
Comparación LDA, QDA y K-NN

Precisión LDA con datos de entrenamiento: 68%
Precisión LDA con datos de prueba: 45%

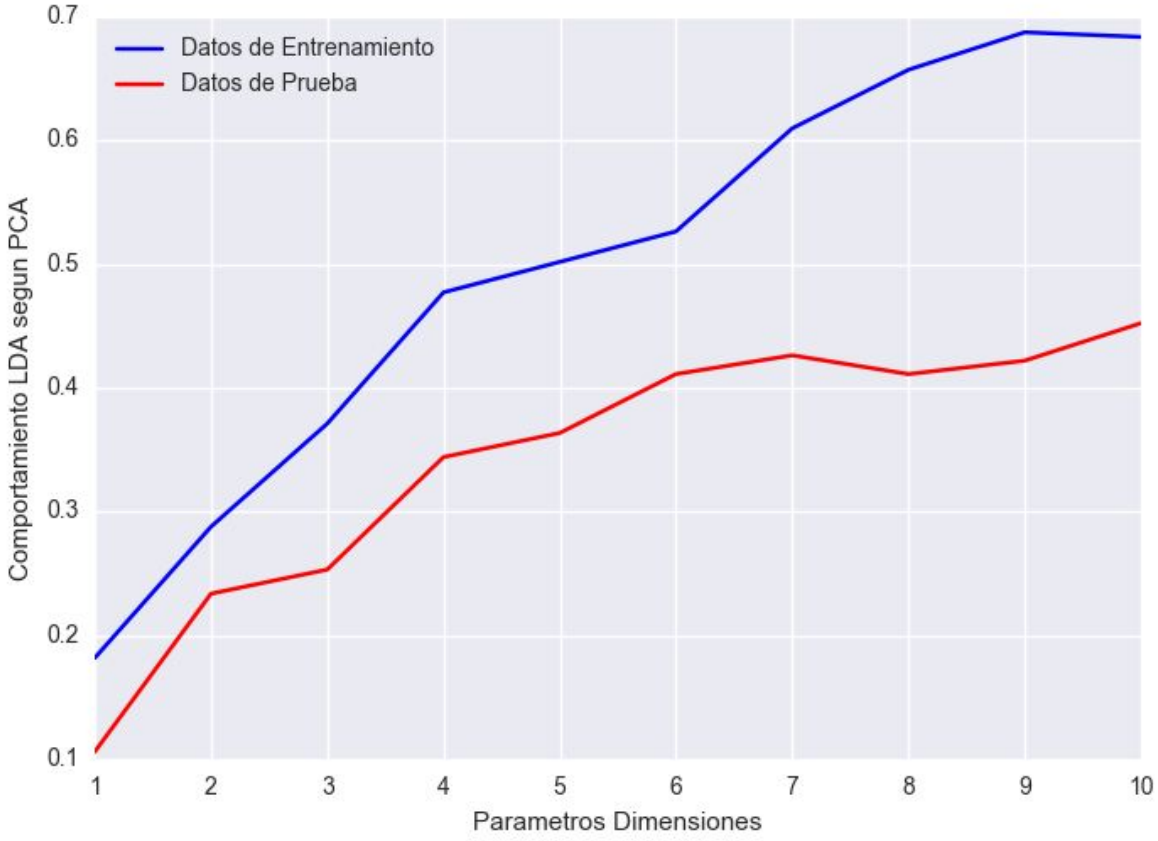
Precisión QDA con datos de entrenamiento: 98%
Precisión QDA con datos de prueba: 42%

Precisión K-NN con datos de entrenamiento: 93%
Precisión K-NN con datos de prueba: 49%

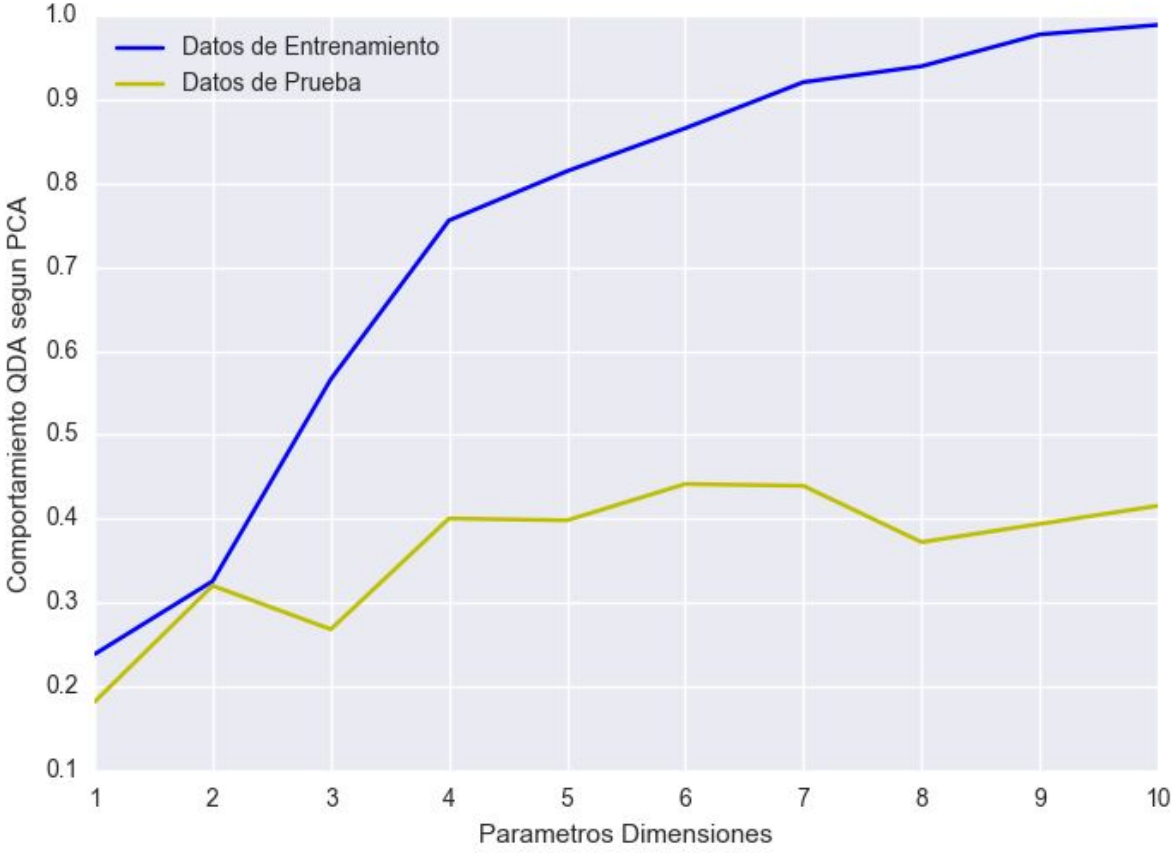
Desempeño K-nn



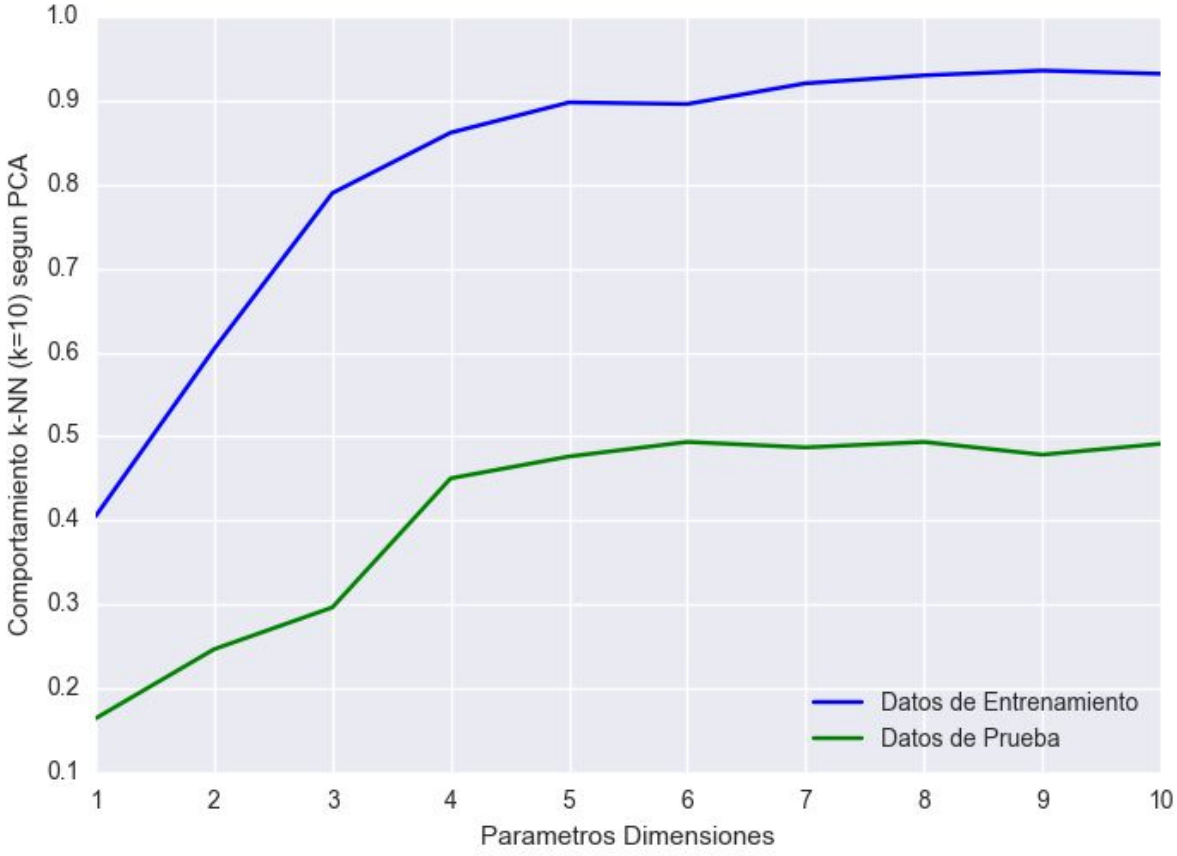
Desempeño de LDA, utilizando PCA



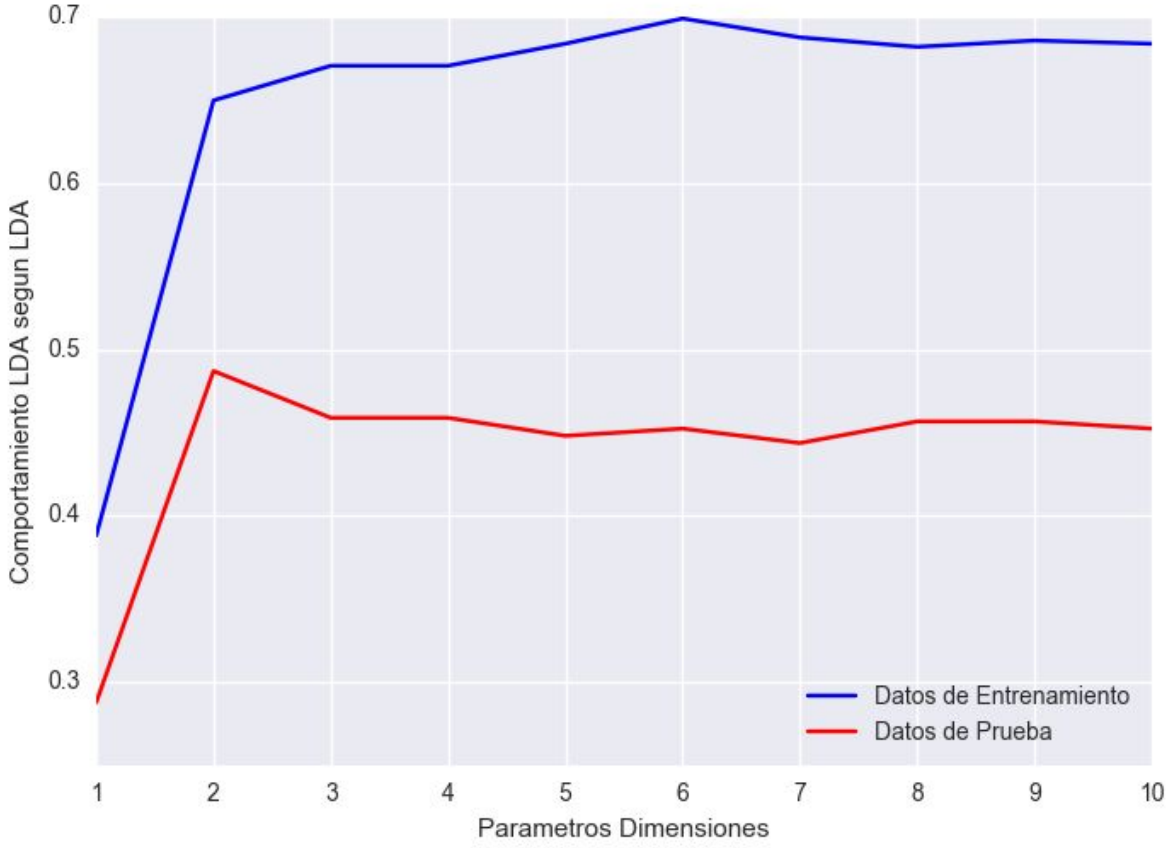
Desempeño de QDA, utilizando PCA



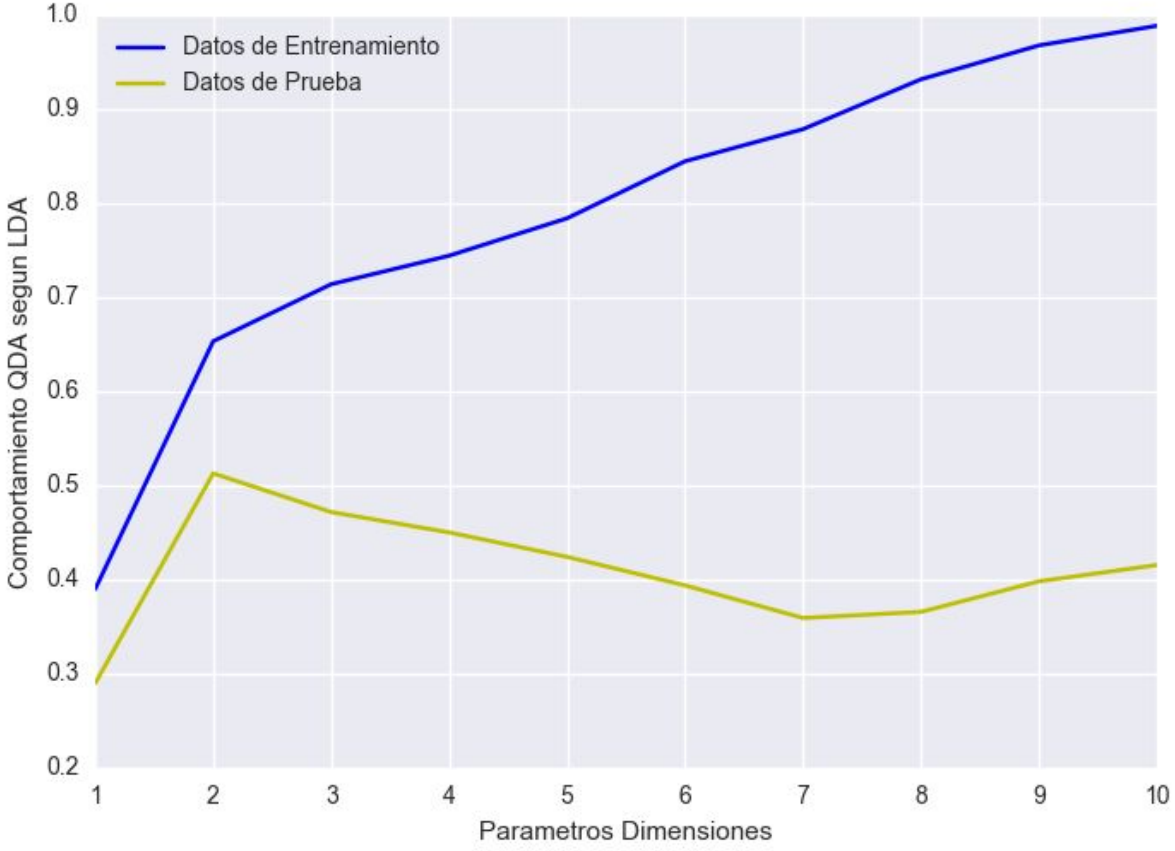
Desempeño de K-nn, utilizando PCA



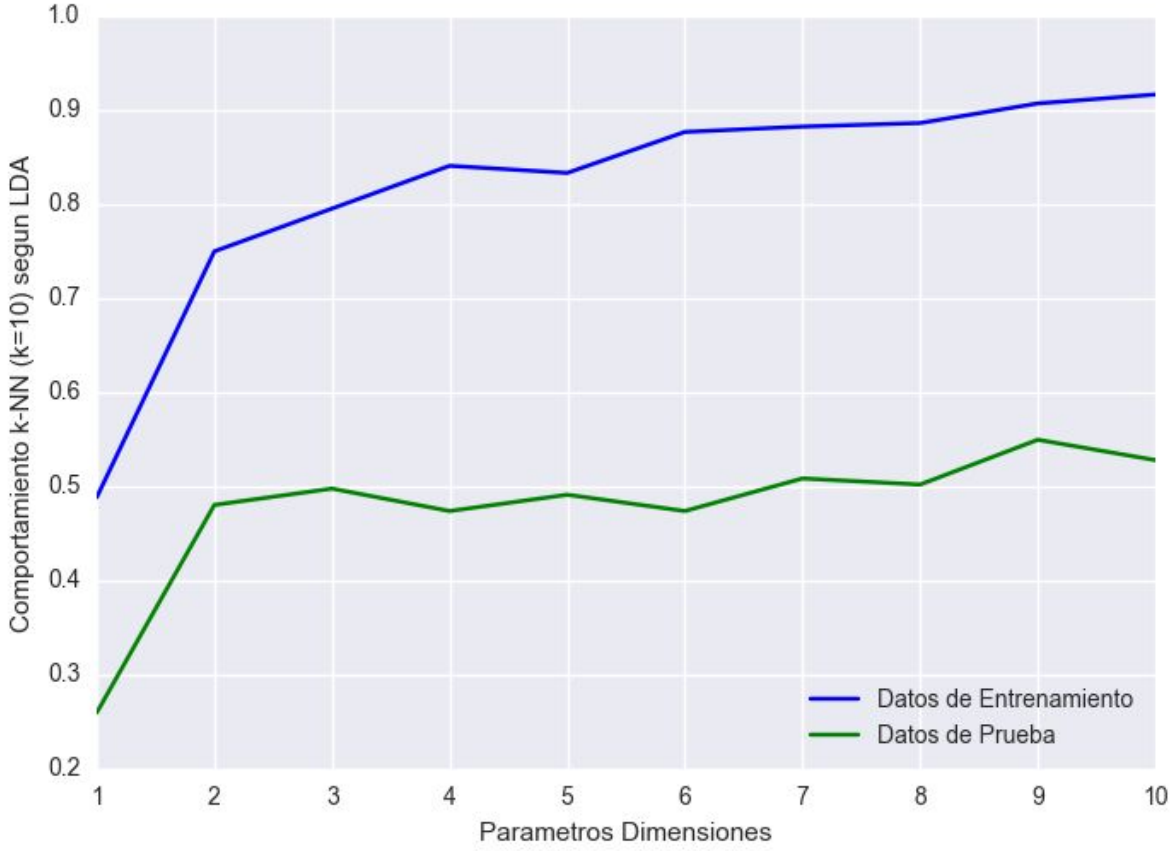
Desempeño de LDA, utilizando Fisher Discriminant



Desempeño de QDA, utilizando Fisher Discriminant



Desempeño de K-nn, utilizando Fisher Discriminant



Análisis de opiniones sobre películas

Preprocesamiento de texto: Stemming v/s lematización

- Se obtienen mejores resultados con lematización: Reduce a su tronco léxico base sólo tokens que corresponden.

- Stemming reduce tokens que no son verbos, como *something* y *already*.

- Ejemplo: *If you're good at something, never do it for free.*
  Stemming: *'re good someth , never free*
  Lematización: *'re good something, never fee*

Descripción de datos

- Set de entrenamiento y de prueba: 3554 registros cada uno.

- Cada registro posee dos características: Text (opinión del usuario) y sentimiento (polaridad de la opinión).

- Polaridad de la opinión: +1 si es positiva, -1 si es negativa.

Construcción de vocabulario

- |Vocabulario| = 9811 tokens.

- Top-ten de tokens: token v/s frecuencia:
  (film, 581)
  (movie, 567)
  (one, 259)
  (ha, 246)
  (like, 239)
  (story, 204)
  (character, 178)
  (time, 176)
  (make, 167)

Desempeño de clasificadores: Clasificador Bayesiano Ingenuo

Las mejores métricas se obtienen al no filtrar stopwords y usar lematización.

Precisión sobre datos de entrenamiento: 0,949
Precisión sobre datos de prueba: 0,757

Table 4: Caso 2: Sin filtrar stopwords y usando lematización

	Precisión	Recall	Valor-F	Soporte
+1	0,74	0,80	0,77	1784
-1	0,78	0,72	0,75	1770
Promedio / Total	**0,76**	**0,76**	**0,76**	**3554**

Ejemplos

- *’a’ for creativity but comes across more as a sketch for a full-length comedy .*
  Probabilidad -1: 0,96
  Probabilidad +1: 0,04

- *it just goes to show , an intelligent person isn’t necessarily an admirable storyteller .*
  Probabilidad -1: 0,53
  Probabilidad +1: 0,47

Desempeño de clasificadores: Clasificador Ingenuo Multinomial

Las mejores métricas se obtienen al no filtrar stopwords y usar lematización.

Precisión sobre datos de entrenamiento: 0,949
Precisión sobre datos de prueba: 0,757

Table 7: Caso 2: Sin filtrar stopwords y usando stemming

	Precisión	**Recall**	**Valor-F**	**Soporte**
+1	0,75	0,78	0,76	1784
-1	0,77	0,74	0,75	1770
Promedio / Total	**0,76**	**0,76**	**0,76**	**3554**

Ejemplos

- *rice never clearly defines his characters or gives us a reason to care about them .*
  Probabilidad -1: 0,94
  Probabilidad +1: 0,06

- *an eccentric little comic/thriller deeply in love with its own quirky personality .*
  Probabilidad -1: 0,23
  Probabilidad +1: 0,77

Desempeño de clasificadores: Regresión Logística Regularizado

- Las mejores métricas se obtienen al no filtrar stopwords y usar lematización.
- Parámetro C: Regularizar, es decir, minimizar error de predicción. Mejores resultados se obtienen para C = 10.

Table 11: Caso 2: Sin filtrar stopwords y usando lematización. C = 10

	Precisión	**Recall**	**Valor-F**	**Soporte**
+1	0,70	0,72	0,71	1784
-1	0,71	0,69	0,70	1770
Promedio / Total	**0,71**	**0,71**	**0,71**	**3554**

Precisión sobre datos de entrenamiento: 1

Precisión sobre datos de prueba: 0,728

Ejemplos

- *entertains by providing good , lively company .*
  Probabilidad -1: 0,01
  Probabilidad +1: 0,99

- *k 19 stays afloat as decent drama/action flick*
  Probabilidad -1: 0,66
  Probabilidad +1: 0,34

Desempeño de clasificadores: SVM Lineal

- Las mejores métricas se obtienen al filtrar stopwords y usar stemming.
- Parámetro C: Evitar que modelo se sobreajuste. Mejores resultados se obtienen para C = 0,1.

Table 16: Caso 3: Filtrando stopwords y usando stemming. C = 0,1

	Precisión	**Recall**	**Valor-F**	**Soporte**
+1	0,72	0,76	0,74	1784
-1	0,74	0,70	0,72	1770
Promedio / Total	**0,73**	**0,73**	**0,73**	**3554**

Precisión sobre datos de entrenamiento: 0,922

Precisión sobre datos de prueba: 0,729

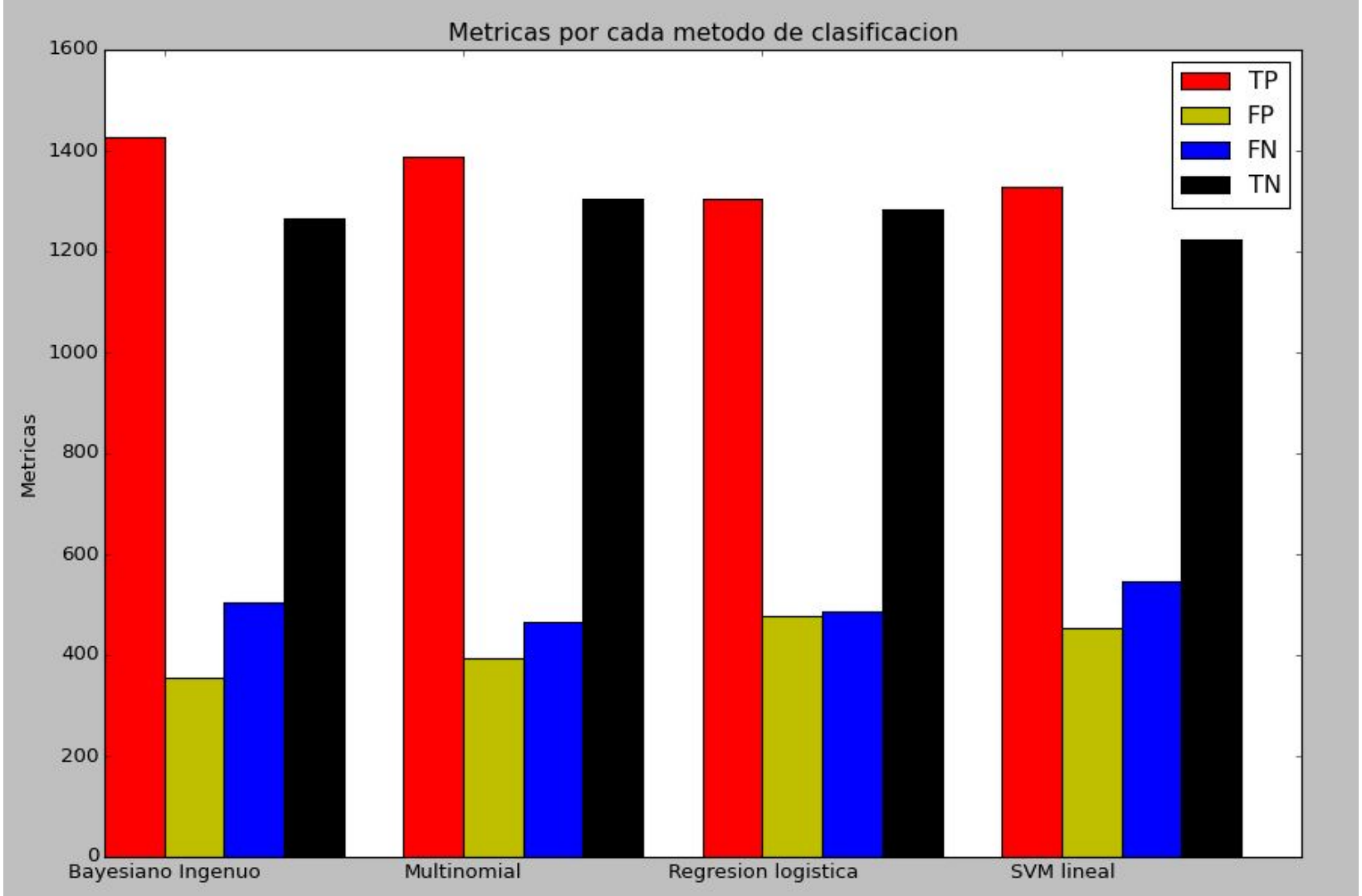
Ejemplos

- *this movie is maddening . it conveys a simple message in a visual style that is willfully overwrought .*
  Probabilidad -1: 0,47
  Probabilidad +1: 0,53

- *what a bewilderingly brilliant and entertaining movie this is*
  Probabilidad -1: 0,14
  Probabilidad +1: 0,86

Comparación de métodos de clasificación

- Se evalúa cada método en base a la cantidad de clasificaciones True Positive (TP), False Positive (FP), True Negative (TN) y False Negative (FN).
- 'el positivo' es la clase +1, y 'el negativo' es la clase -1.
- Mejor resultado clase +1 (TP): Modelo Bayesiano Ingenuo (ver gráfico).
- Mejor resultado clase -1 (TN): Modelo Ingenuo Multinomial (ver gráfico).


Metricas por cada metodo de clasificacion
TP
FP
FN
TN
Metricas
1600
1400
1200
1000
800
600
400
200
0
Bayesiano Ingenuo
Multinomial
Regresion logistica
SVM lineal

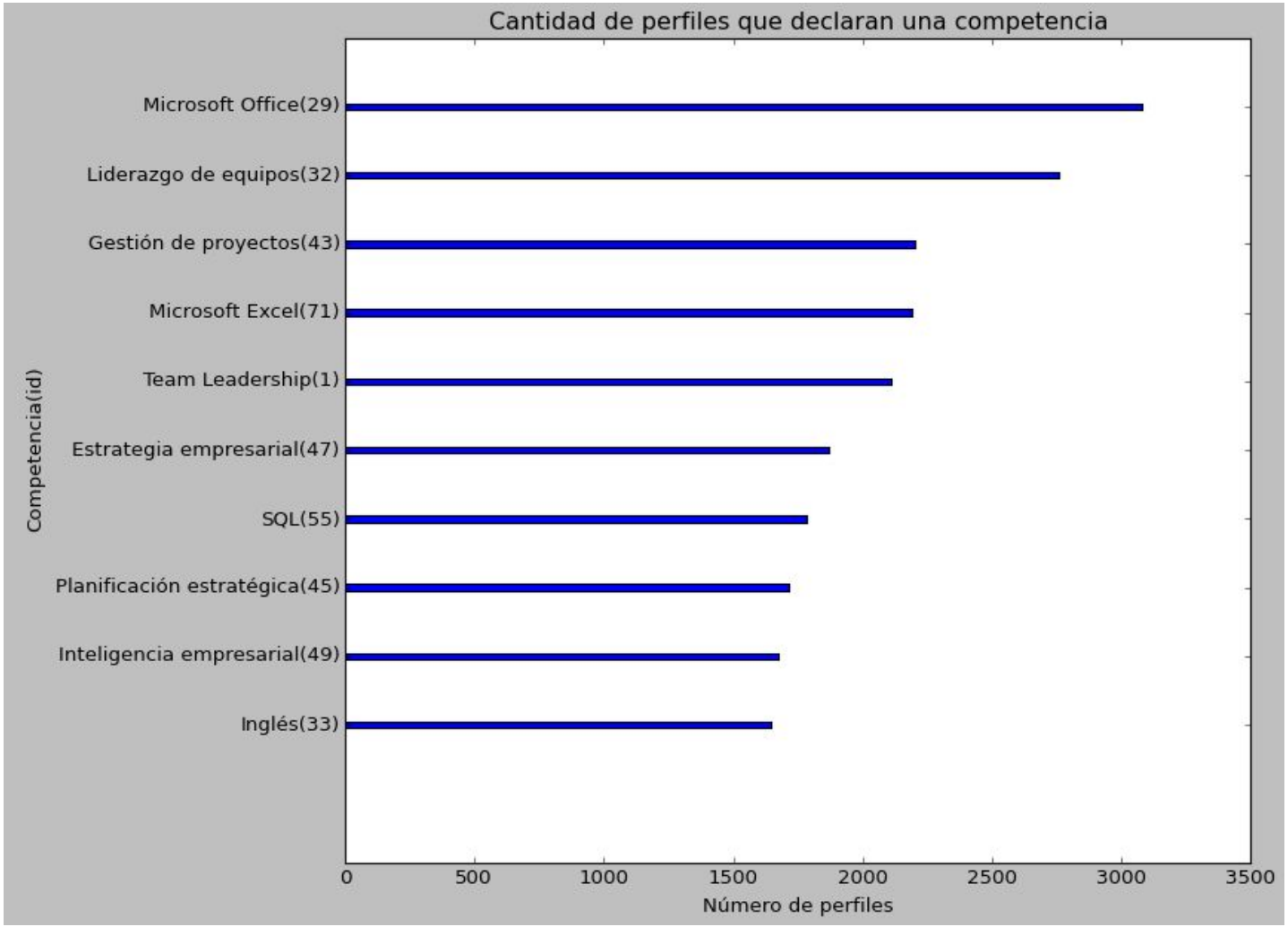
Skill Prediction en LinkedIN

Descripción de dataset

- Existen 7890 perfiles y 14544 competencias.
- Datos se almacenan en matriz sparse Z, de dimensiones 7890 x 14544.
- La casilla ij de Z posee el valor 1 si el perfil i declara la competencia j. En caso contrario, su valor es 0.
- Conjunto de entrenamiento: Corresponde al 60% de los registros de Z.
- Conjunto de prueba: Corresponde al 40% de los registros de Z.

Descripción de dataset

- Existen 7890 perfiles y 14544 competencias.
- Datos se almacenan en matriz sparse Z, de dimensiones 7890 x 14544.
- La casilla ij de Z posee el valor 1 si el perfil i declara la competencia j. En caso contrario, su valor es 0.
- Conjunto de entrenamiento: Corresponde al 60% de los registros de Z.
- Conjunto de prueba: Corresponde al 40% de los registros de Z.


Cantidad de perfiles que declaran una competencia
Microsoft Office(29)
Liderazgo de equipos(32)
Gestión de proyectos(43)
Microsoft Excel(71)
Team Leadership(1)
Estrategia empresarial(47)
SQL(55)
Planificación estratégica(45)
Inteligencia empresarial(49)
Inglés(33)
Competencia(id)
0
500
1000
1500
2000
2500
3000
3500
Número de perfiles

Predicción de la competencia *Microsoft Office*

- Corresponde a la competencia más frecuente.

- Se entrenan y evalúan cinco tipos de clasificadores:
  - Bayesiano Ingenuo Binario
  - Bayesiano Ingenuo Multinomial
  - Regresión Logística Regularizado
  - SVM Lineal
  - Vecinos Más Cercanos (k-NN)

- Existen dos clases:
  - Clase 1: Competencia es declarada en perfil.
  - Clase 0: Caso contrario.

Análisis de resultados

- Regresión logística: Precisión sobre conjunto de pruebas obtiene su mejor resultado para C = 0.01. Para siguientes instancias, precisión disminuye. SVM presenta mismo comportamiento.

- Modelos binario y multinomial presentan buenos resultados, mas no mejores que regresión logística y SVM (resultado válido para siguientes predicciones).

- k-NN presenta los peores resultados. K = 1 es el mejor.

- Mejor resultado: Regresión logística, C = 0.01
  - Precisión de entrenamiento: 0.843684
  - Precisión de pruebas: 0.829840

Predicción de la competencia *Liderazgo de equipos*

- En general, valores más altos de precisión (entrenamiento y prueba) para todos los clasificadores respecto a predicción de *Microsoft Office*.

- Resultados similares a predicción anterior, pero regresión logística obtienen valores más altos de precisión para C = 0.1, al igual que SVM.

- Nuevamente, regresión logística es el clasificador más preciso, con C = 0.1.
  - Precisión de entrenamiento: 0.915927
  - Precisión de prueba: 0.895120

Predicción de la competencia *Gestión de proyectos*

- En general, valores inferiores de precisión (entrenamiento y prueba) para todos los clasificadores respecto a predicción de *Liderazgo de equipos*, pero superiores a predicción de *Microsoft Office*.

- Resultados similares a predicciones anteriores, nuevamente con C = 0.1 tanto para regresión logística como SVM.

- Por tercera vez, regresión logística es el clasificador más preciso, con C = 0.1.
  - Precisión de entrenamiento: 0.912759
  - Precisión de prueba: 0.873574