Tarea 2 : Métodos Lineales para Clasificación

Juan Pablo Castillo
Álvaro Rojas

1. Reducción de Dimensionalidad para Clasificación

Datos Utilizados

- Se trabajó con una colección de sonidos fonéticos que deben ser identificados con vocales del inglés británico.
- Los datos son representados en un espacio de d = 10 características. Existen 528 datos de entrenamiento y 462 de pruebas.
- Existen 11 tipos de vocales que serán las etiquetas de las categorías.
- Los experimentos constan de hacer repetir a una persona 6 veces cada vocal, y así para 15 personas obteniéndose los 990 datos que se separan en train y test.

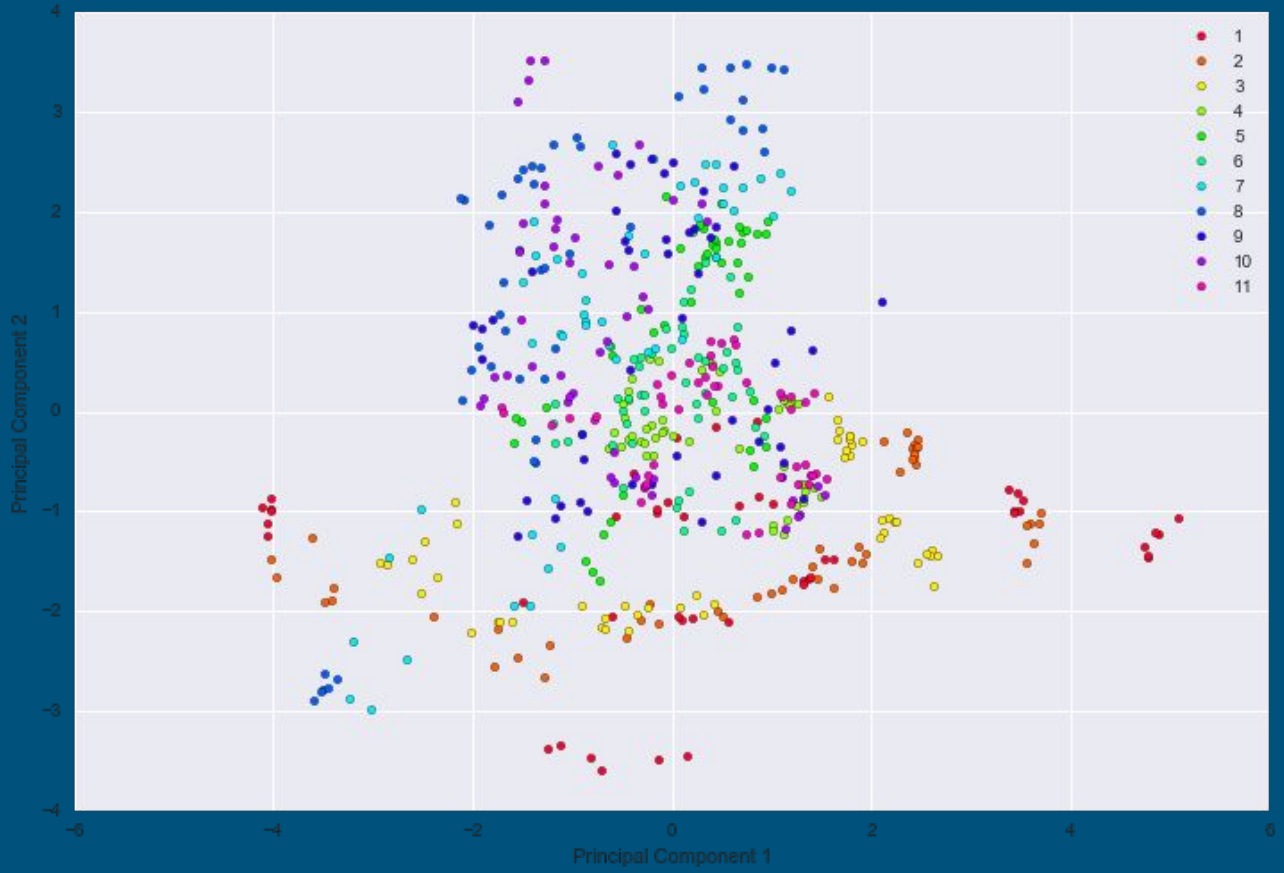
Datos Utilizados

Algunos aspectos importantes de los datos(de entrenamiento).

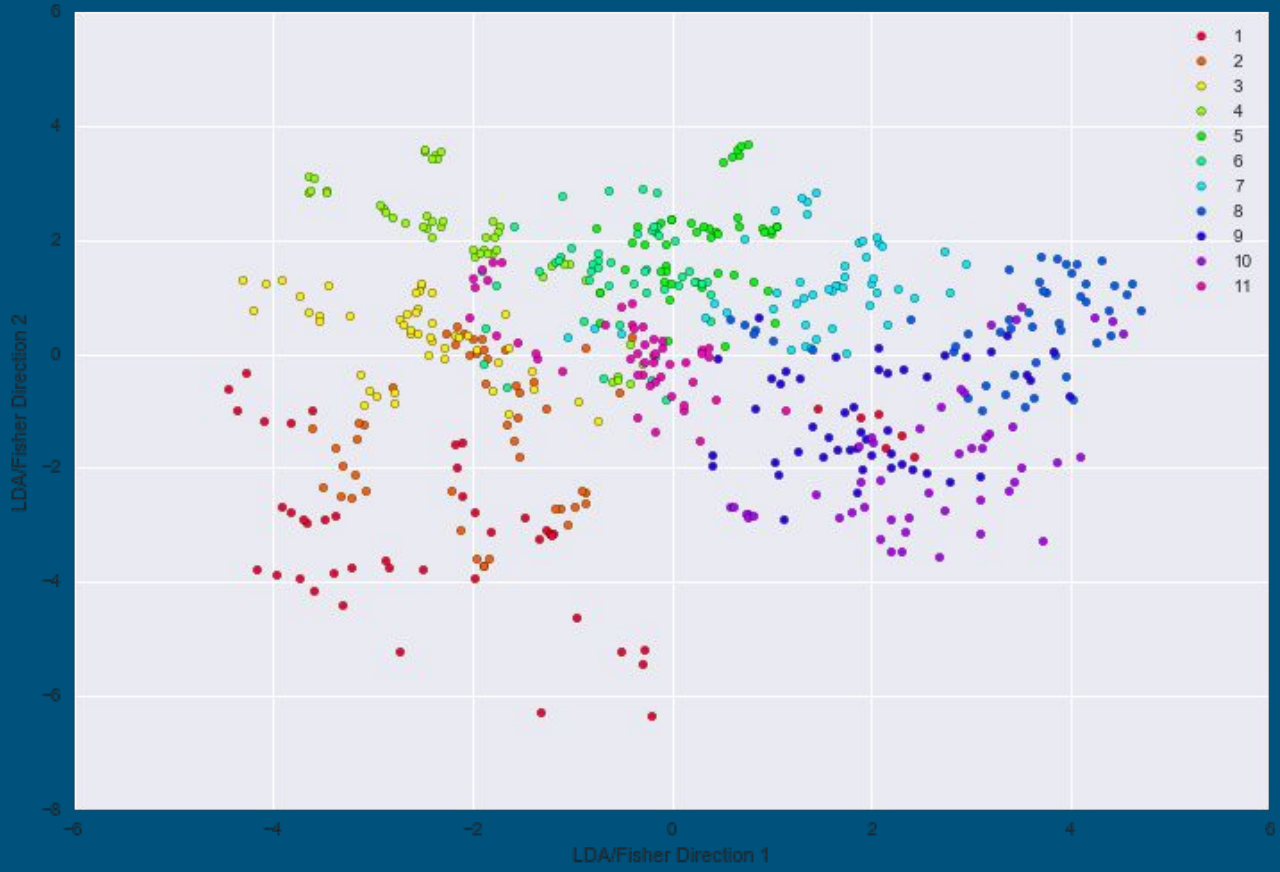
```
y         6.000000
x.1      -3.166695
x.2       1.735343
x.3      -0.448002
x.4       0.524983
x.5      -0.389280
x.6       0.584960
x.7       0.017477
x.8       0.417394
x.9      -0.268112
x.10     -0.084568
dtype: float64
```

	y	x.1	x.2	x.3	x.4	x.5	x.6	x.7	x.8	x.9	x.10
row.names											
524	7	-4.065	2.876	-0.856	-0.221	-0.533	0.232	0.855	0.633	-1.452	0.272
525	8	-4.513	4.265	-1.477	-1.090	0.215	0.829	0.342	0.693	-0.601	-0.056
526	9	-4.651	4.246	-0.823	-0.831	0.666	0.546	-0.300	0.094	-1.343	0.185
527	10	-5.034	4.993	-1.633	-0.285	0.398	0.181	-0.211	-0.508	-0.283	0.304
528	11	-4.261	1.827	-0.482	-0.194	0.731	0.354	-0.478	0.050	-0.112	0.321

Representación PCA



Representación LDA



PCA vs LDA

- La representación de los datos de PCA no obtiene una división o agrupación clara de estos. En cambio en el gráfico de LDA se observa un mayor orden o cercanía entre clases del mismo tipo.
- LDA considera las etiquetas de las clases para calcular los vectores que maximizan la separación entre ellas. Por otro lado, PCA se opta por las direcciones que maximicen la variabilidad.
- Elegir LDA para cuando se tienen datos con etiquetas y PCA en caso contrario.

Clasificador en base a Probabilidad

- Existen la misma cantidad de datos para cada etiqueta lo que provoca una probabilidad similar para todas las clases.
- Entre mayor cantidad de clases, manteniendo la cantidad de datos parecida para todas las etiquetas, menos efectivo es este método.

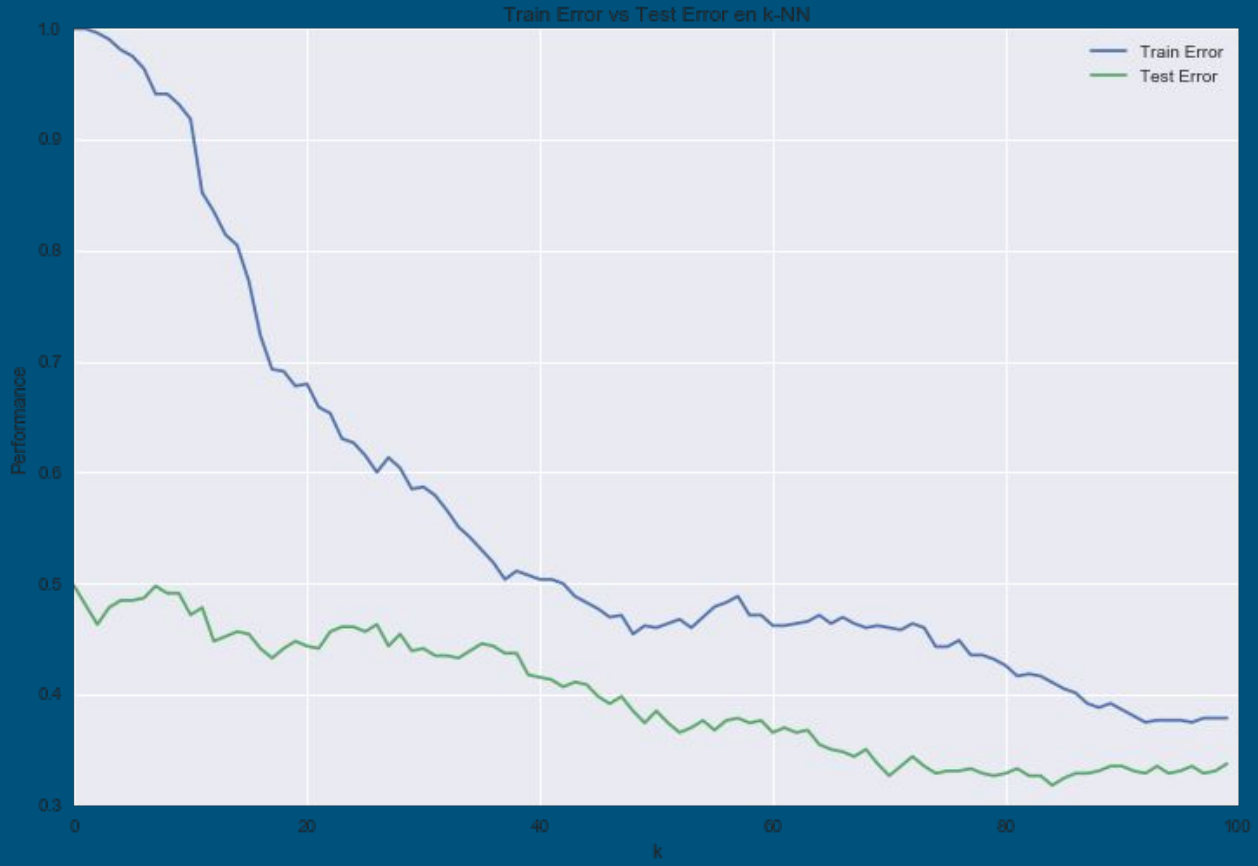
```
Probabilidad Clase 1 = 0.090909
Probabilidad Clase 2 = 0.090909
Probabilidad Clase 3 = 0.090909
Probabilidad Clase 4 = 0.090909
Probabilidad Clase 5 = 0.090909
Probabilidad Clase 6 = 0.090909
Probabilidad Clase 7 = 0.090909
Probabilidad Clase 8 = 0.090909
Probabilidad Clase 9 = 0.090909
Probabilidad Clase 10 = 0.090909
Probabilidad Clase 11 = 0.090909
Clase: 1
```

Comparación LDA, QDA y un modelo k-NN.

- El con mejor error de entrenamiento fue el que peor se comportó en el set de pruebas(sobreajuste).
- Ocurre lo inverso con LDA que, a pesar de tener el peor error en el conjunto de entrenamiento, logró un valor de error más aceptable que QDA.

```
LDA - Conjunto de entrenamiento: 0.683712
LDA - Conjunto de prueba: 0.452381
QDA - Conjunto de entrenamiento: 0.988636
QDA - Conjunto de prueba: 0.415584
KNN - Conjunto de entrenamiento con k=10: 0.931818
KNN - Conjunto de prueba con k=10: 0.491342
```

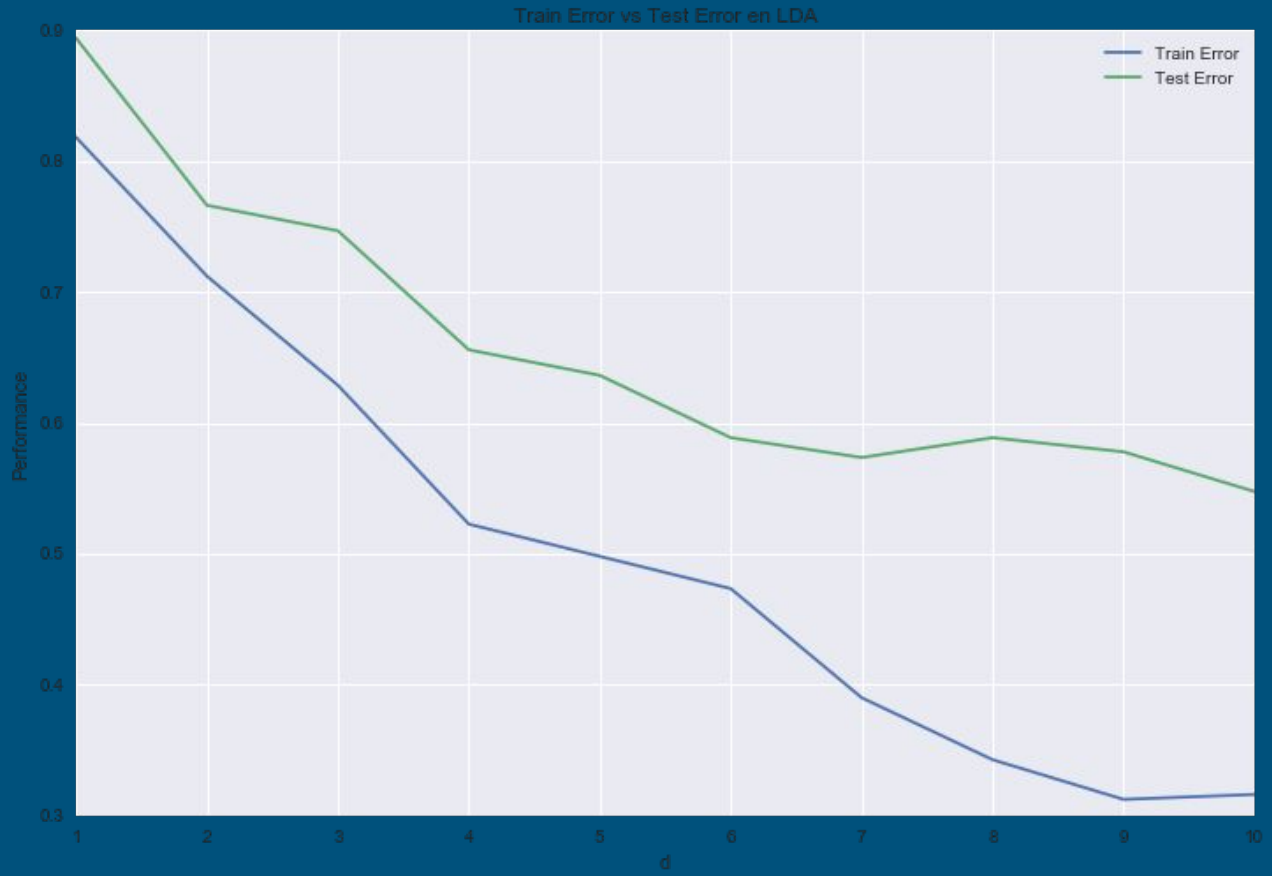
Comparación LDA, QDA y un modelo k-NN.



Comparación LDA, QDA y un modelo k-NN.

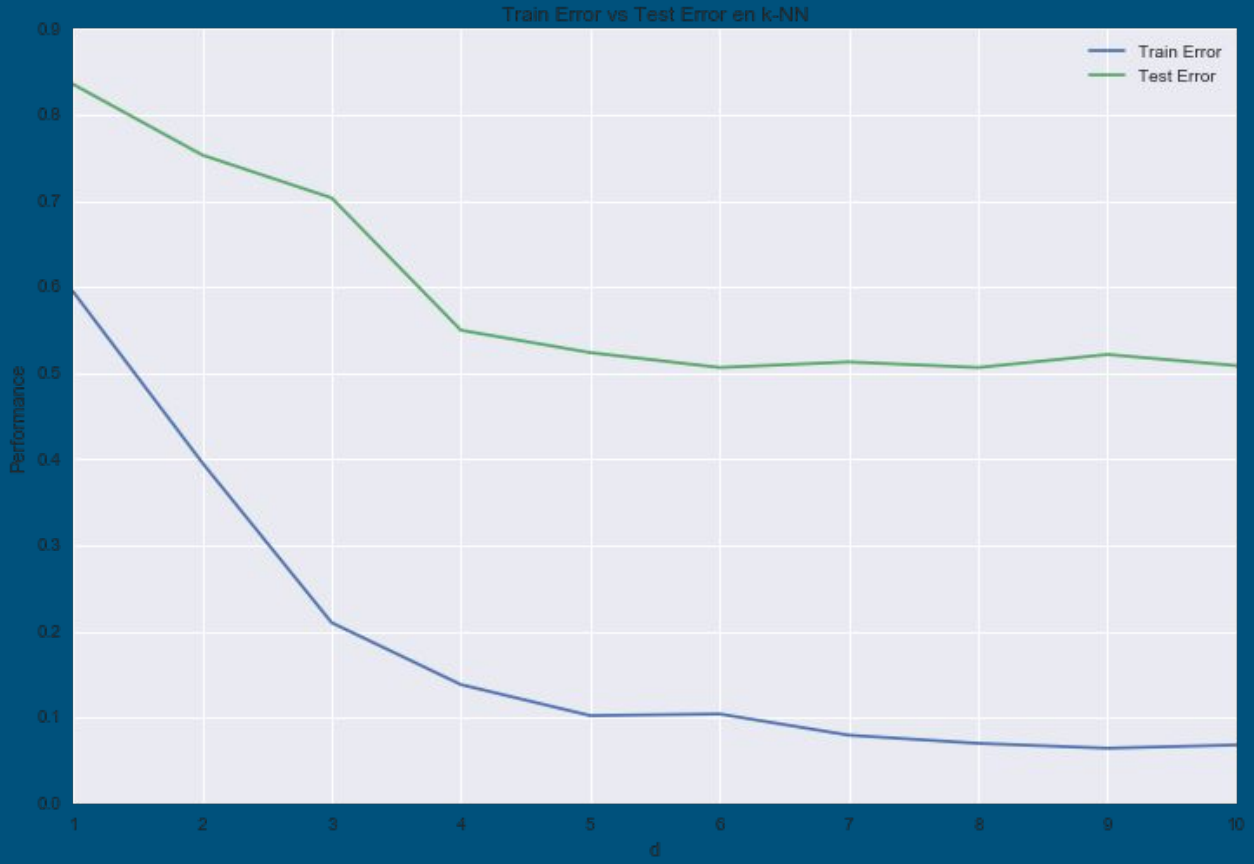
- La elección del mejor k en este caso sería uno pequeño.
- El comportamiento puede deberse a que en general cuando se consideran conjuntos de vecinos pequeños se es más fácil tener un parecido fuerte entre los pocos integrantes.
- La exactitud de este algoritmo puede ser severamente degradada por la presencia de ruido o características irrelevantes.

PCA entrenando con LDA



PCA entrenando con QDA

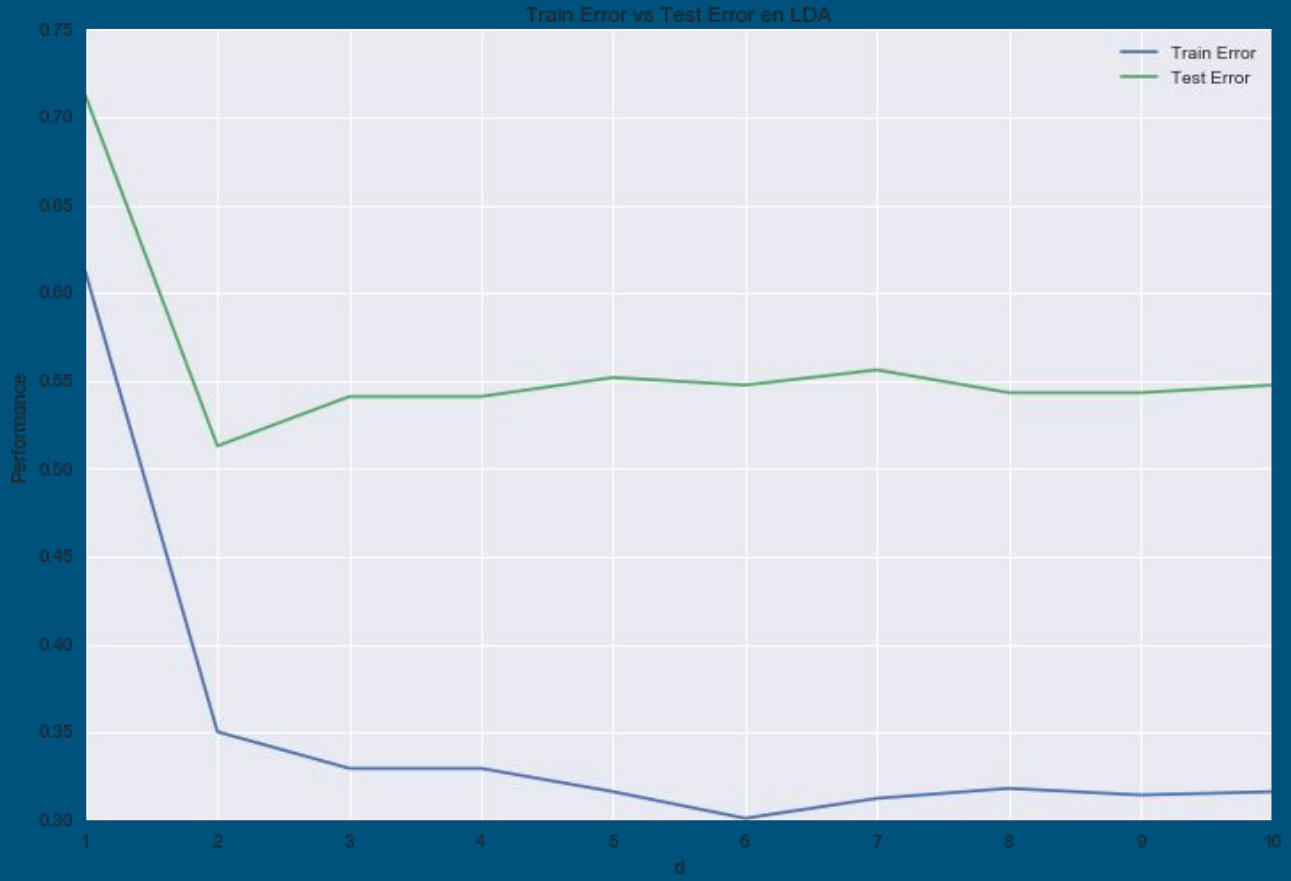
PCA entrenando con k-NN



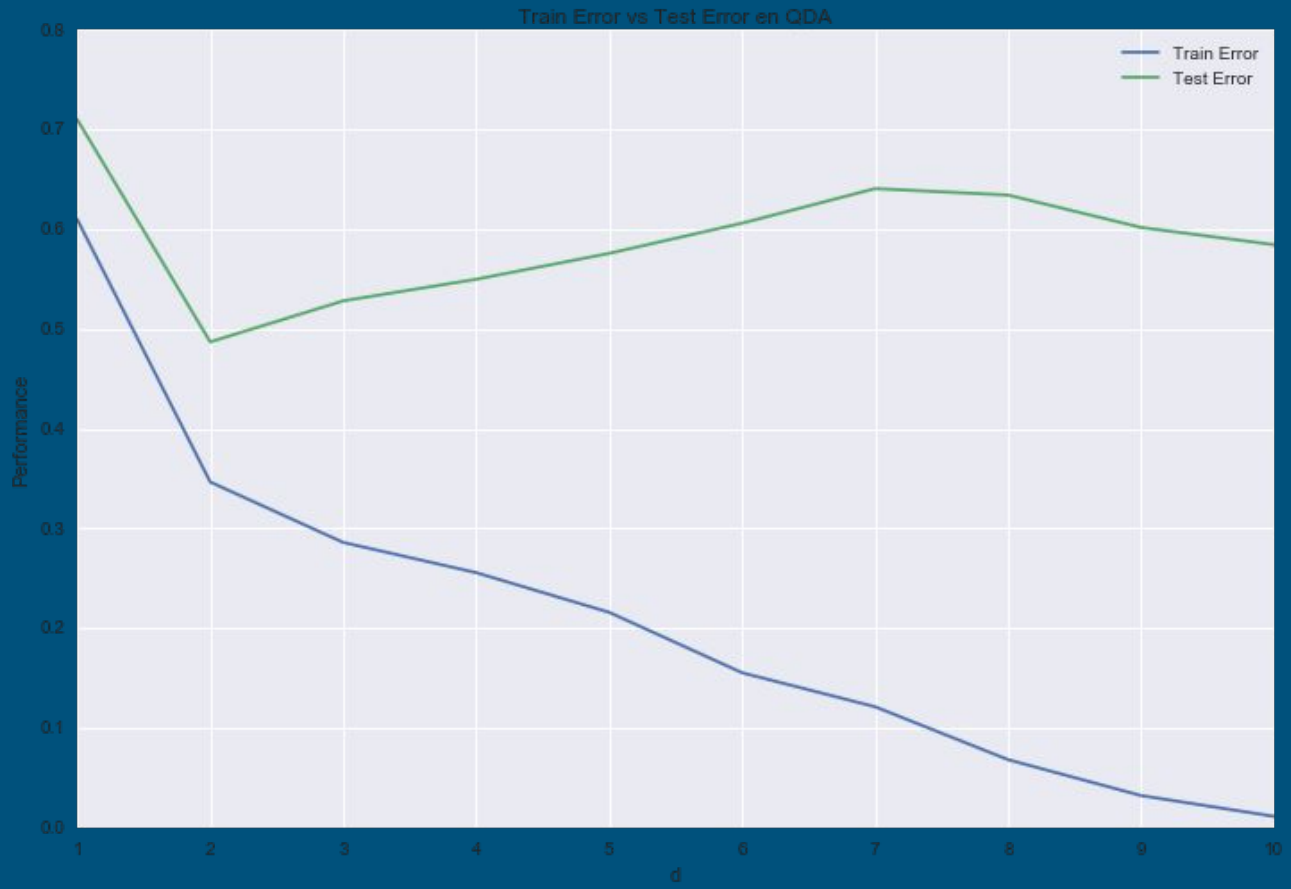
PCA entrenando con LDA, QDA y k-NN.

- El error de entrenamiento disminuye por cada incorporación de una componente.
- El error de entrenamiento es menor que para cada valor de dimensionalidad.
- Los valores de dimensión en los cuales los métodos alcanzan su mínimo error de prueba son: 10 para LDA (0.547619), 6 para QDA (0.558442) y 6 para k-NN (0.506494).
- Se puede observar una leve mejoría en la mayoría de los casos a lo obtenido sin PCA, pero no significantes.

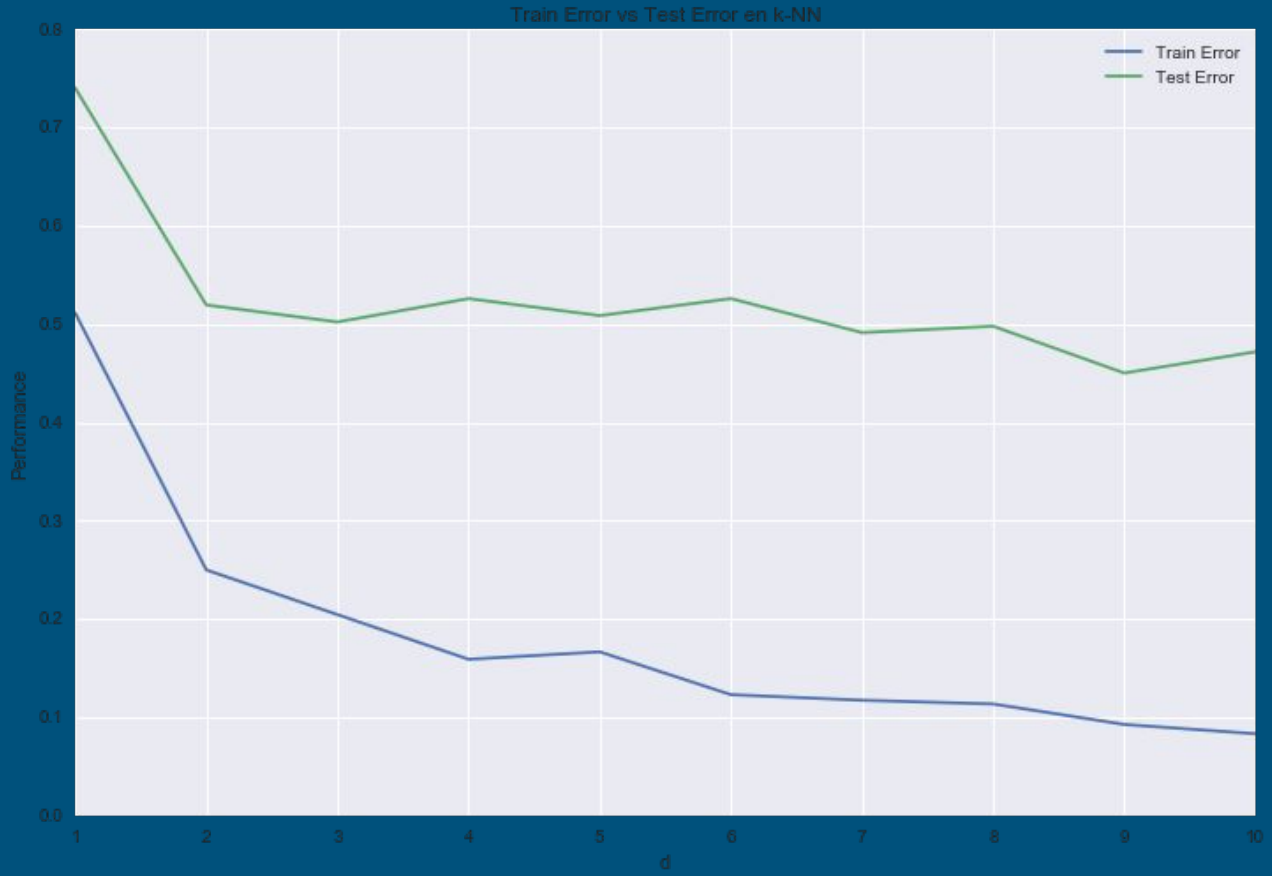
LDA entrenando con LDA



LDA entrenando con QDA



LDA entrenando con k-NN



LDA entrenando con LDA, QDA y k-NN.

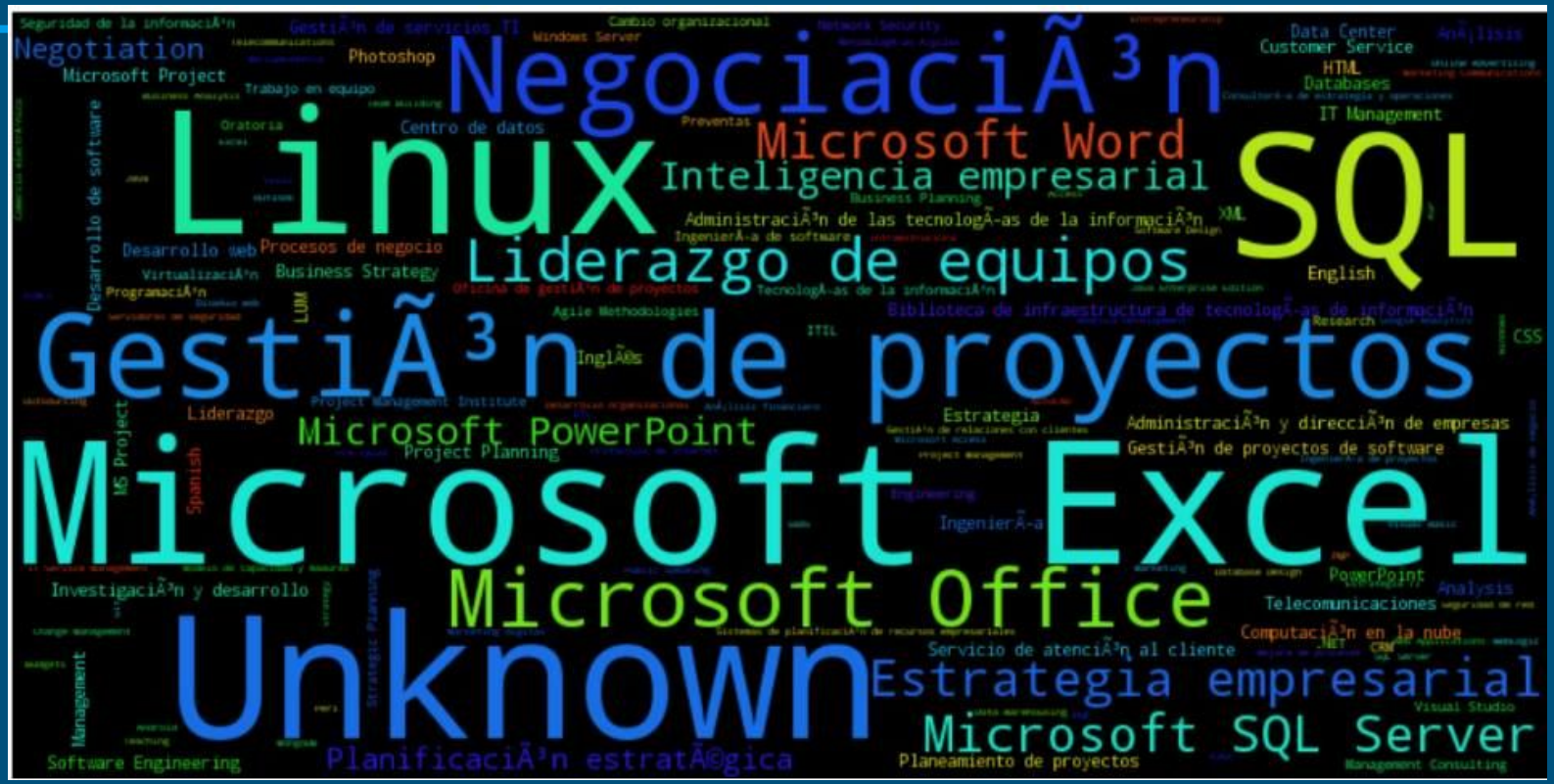
- El error de entrenamiento por lo general disminuye por cada incorporación de una componente.
- Los valores de dimensión en los cuales los métodos alcanzan su mínimo error de prueba son: 2 para LDA (0.512987), 2 para QDA (0.487013) y 9 para k-NN (0.450216).
- Los errores son bastantes cercanos a los valores obtenidos sin haber realizado LDA.
- Estos se consiguen en la mayoría de los casos con un valor muy bajo de componentes disminuyendo complejidad y cómputo.
- Se confirma que LDA es una bueno para disminuir dimensionalidad en problemas con etiquetas conocidas a diferencia de PCA.

3. Skill Prediction en LinkedIN

Construyendo y Analizando Matriz Dispersa

- Se genera una matriz dispersa 7890x14544 en donde el usuario i contiene la habilidad j se asigna un 1, en caso contrario un 0. Las habilidades varían entre 0 y 14544, para un número de 7890 usuarios.
- Se divide la matriz en dos sub-matrices, de entrenamiento y prueba.
- Se agregaron de forma aleatória 70% de los datos de la matriz original a la matriz de entrenamiento y el 30% restante a la matriz de prueba, obteniendo como resultado las matrices de tamaños (5523, 14544) y (2367, 14544).

Gráfica de Nube de Palabras



Gráfica de Nube de Palabras

- Las palabras con mayores tamaños, y que por ende son las con mayor frecuencia en las personas que dicen tener esa habilidad, son Microsoft Excel, Linux, SQL, Negociación, gestión de proyectos y "unknown"(variable sin nombre en los datos que se corrigió manualmente).
- Es posible presumir que la mayoría los usuarios de LinkedIN son personas que deberían poder trabajar en cualquier “empresa clásica”, dirigiendo y organizando personal además de poder realizar algunos trabajos menores de secretaría y finanzas.

Matrices para predicción de competencias

- Se escoge como objetivo la skill "Microsoft Excel" cuya id es 71. Recordar que esta variable parecía ser la que más se repetía cuando se realizó el gráfico de competencias entre todos los usuarios, por lo que es probable que siga estando en los primeros lugares dentro del set de entrenamiento.

Entrenamiento de clasificadores