

# Estudio comparativo de métodos espectrales para reducción de la dimensionalidad: LDA versus PCA

Comparative study between spectral methods for dimension reduction  
LDA versus PCA

**Andrés J. Anaya-Isaza**  
**Diego H. Peluffo-Ordoñez**  
**Juan C. Alvarado-Pérez**  
Universidad Surcolombiana,  
Universidad Tecnológica de Pereira  
Universidad Técnica del Norte,  
Ibarra, Ecuador  
Corporación Universitaria Autónoma  
de Nariño Pasto, Colombia  
[andres.anaya@usco.edu.co](mailto:andres.anaya@usco.edu.co)  
[dhpeluffo@utn.edu.ec](mailto:dhpeluffo@utn.edu.ec)  
[jcalvarado@usal.es](mailto:jcalvarado@usal.es)

**Jorge Ivan-Rios**  
**Juan A. Castro-Silva**  
**Paul D. Rosero-Montalvo**  
Universidad Tecnológica de Pereira  
Universidad Surcolombiana de  
Pereira - Risaralda y Neiva-Huila,  
Colombia  
Universidad Técnica del Norte,  
Ibarra, Ecuador  
[juancasi@usco.edu.co](mailto:juancasi@usco.edu.co)  
[jirios@utp.edu.co](mailto:jirios@utp.edu.co)  
[pdrosero@utn.edu.ec](mailto:pdrosero@utn.edu.ec)

**Diego F. Peña-Unigarro**  
**Jose A. Salazar-Castro**  
**Juan C. Ana C. Umaquinga-Criollo**  
Universidad de Nariño-Pasto  
Colombia  
Universidad Nacional de Colombia  
sede-Manizales Colombia  
Universidad Técnica del Norte,  
Ibarra, Ecuador  
[diferpun@gmail.com](mailto:diferpun@gmail.com)  
[alejo26st@udenar.edu.co](mailto:alejo26st@udenar.edu.co)  
[anacristinaucr@gmail.com](mailto:anacristinaucr@gmail.com)

**Resumen**— Este trabajo presenta un estudio comparativo con métodos de reducción de la dimensión lineal, tales como: Análisis de Componentes Principales & Análisis Discriminante Lineal. El estudio pretende determinar, bajo criterios de objetividad, cuál de estas técnicas obtiene el mejor resultado de separabilidad entre clases. Para la validación experimental se utilizan dos bases de datos, del repositorio científico (UC Irvine Machine Learning Repository), para dar tratamiento a los atributos del data-set en función de confirmar visualmente la calidad de los resultados obtenidos. Las immersiones obtenidas son analizadas, para realizar una comparación de resultados del embejimiento representados con  $RNX(K)$ , que permite evaluar el área bajo la curva, del cual asume una mejor representación en una topología global o local que posteriormente genera los gráficos de visualización en un espacio de menor dimensión, para observar la separabilidad entre clases conservando la estructura global de los datos.

**Palabras clave**— Análisis de componentes principales, Análisis discriminante lineal, Aprendizaje de máquina, Clasificación lineal, Clasificación supervisada, Métodos de reducción de la dimensión.

**Abstract**— This paper presents a comparative study with methods of reducing the linear dimension, such as Principal Component Analysis & linear discriminant analysis. The study aims to determine, based on criteria of objectivity, which of these techniques get the best result of separability between classes. For experimental validation two databases, scientific repository (UC Irvine Machine Learning Repository) are used to provide treatment to the attributes of the data-set according to visually confirm the quality of the results. The dives obtained are analyzed, for comparison of results embejimiento represented with  $RNX(K)$ , which evaluates the area under the curve, which assumes better representation in a local global topology or subsequently generates graphics display a space of smaller dimension, to

observe the class separability preserving the overall structure of the data.

**Keywords** — Principal component analysis, linear discriminant analysis, machine learning, linear classification, supervised classification, methods for dimensionality reduction.

## I. INTRODUCCIÓN

El termino reducción de la dimensión (RD), es el proceso de representar una proyección de muestras, a un espacio de poca dimensionalidad proveniente de una alta dimensión de datos, el cual promete preservar la mayor información intrínseca de las muestras originales [1], esta nueva proyección en caso de satisfacer la tarea de compactar la información conservando la mayor parte de estos datos, conviene en labores de aprendizaje supervisado, que concede realizar tareas de clasificación y visualización [2], donde cada muestra posee etiquetas como elementos descriptivos que representan las diferentes clases, Análisis discriminante lineal bajo el criterio de Fisher (LDA) [3] pertenece al conjunto de métodos más usados de la literatura de RD. Este método busca que las muestras para cada uno de los datos que siguen patrones de distribución gaussianas, obtengan resultados que comprometa una transformación embebida para garantizar la máxima dispersión entre clases para reducir a su mínimo posible la dispersión interna de cada clase [3]. Análisis de componentes principales (PCA) es el segundo método a confrontar, siendo popular para tareas de reducción de la dimensión, así como también en tareas como

aprendizaje de máquina y reconocimiento de patrones. Este método especialmente obtiene buenos resultados en selección y extracción de características, de igual manera que (LDA) [4], PCA se caracteriza por proyectar una transformación lineal que expresa un nuevo sistema de vectores ortonormales, el cual describe el conjunto original de datos, este proceso consta de maximizar la varianza para capturar el mayor número de muestras del conjunto de información y almacenarlos en nuevos ejes, donde el primer componente principal, es aquel eje de mayor varianza, el segundo componente principal posee la segunda varianza más grande, y así sucesivamente [5].

El problema a tratar de este trabajo es la comparación cuantitativa de dos métodos espectrales para la reducción de la dimensión tales como: (LDA) & (PCA), mediante la proyección de los datos finales dada en la transformación de la información de cada uno de estos, resolviendo el clásico problema de valores propios para la matrices del proceso a reducir y saber que método realiza mejor su trabajo para tareas de clasificación y compresión de datos [6]. Para comparar los métodos de reducción de la dimensión, se utiliza dos bases de datos del repositorio científico UCI – Machine Learning, con características en sus variables de tipo discreto y continuo, de dimensión inicial 13 para el primer conjunto de datos y de dimensión inicial 9 para el segundo. Posteriormente se aplican cada uno de los algoritmos mediante la implementación del toolbox, Reduction Dimensionality for MatLab [7], por medio de la herramienta científica MATLAB [8].

En la siguiente sesión se (Base Teórica), se profundiza sobre el proceso del desarrollo para estos dos métodos (LDA) & (PCA), para comprender su funcionamiento desde la teoría, el cual conlleva a sus posteriores implementaciones para cada uno de los conjuntos de datos a tratar.

## II. BASE TEÓRICA

### A) PCA

PCA o Principal Component Analysis, es un método espectral de análisis multivariado, basado en correlaciones entre variables, mediante un proceso estocástico que satisface la reducción de dimensiones, realizando una transformación lineal de las variables iniciales, para proyectar vectores propios ortonormales denominados componentes principales [9].

Se tiene un conjunto de datos, el cual se denomina  $\mathbf{X}$ . Se saca el promedio de  $\mathbf{X}$  y se resta para cada coeficiente, en el conjunto de entrenamiento  $\mathbf{X} - \bar{\mathbf{X}}$ . Obteniendo el conjunto de datos:

$$1) \mathbf{X} = \mathbf{X} - \bar{\mathbf{X}} = \mathbf{X} - \bar{\mathbf{X}}$$

Posteriormente se crea una matriz que podemos llamar  $\mathbf{X}$  tal que, cada uno de los elementos de la columna serán datos de

muestra  $\mathbf{x}_i$ , la matriz de nuestro conjunto de datos que representa la matriz de covarianza, para poder determinar los nuevos ejes o vectores ortonormales que componen los componentes principales inmersos en la matriz de covarianza [10], se desarrollan de la siguiente manera:

$$2) \mathbf{C} = \frac{1}{N} \mathbf{X}^T \mathbf{X}; \mathbf{C} \mathbf{E} = \mathbf{E} \mathbf{\Lambda}$$

$\mathbf{\Lambda}$  es la matriz diagonal de valores propios y  $\mathbf{E}$  es la matriz de vectores propios ortonormales.

El resultado de  $\mathbf{E}$  es un nuevo vector o matriz, que transforma la base para conseguir nuevos vectores el cual rotan de posición a sus ejes antecesores para ser almacenados en los vectores propios, donde el primer componente principal, es aquel eje de mayor varianza, el segundo componente principal posee la segunda varianza más grande, y así sucesivamente. Se esperan que en la menor cantidad de ejes se recoja toda la información [11].

### B) Linear Discriminant Analysis (LDA)

Análisis discriminante lineal bajo el criterio de Fisher, este método busca las muestras que sigan patrones de distribución gaussianas para un buen funcionamiento. Dentro de sus aplicaciones más usadas en la literatura, se recomienda para tareas de aprendizaje supervisado & clasificación de clases, ya que proyecta los datos a una baja dimensión en comparación de los datos originales, pero garantiza la máxima dispersión entre clases para reducir a su mínimo posible la dispersión interna de cada clase [3]. La tarea es proyectar el conjunto de datos a una baja dimensión mediante un vector  $\mathbf{S}$  que pueda garantizar el máximo en dispersión o separación de las clases [1] [11]. Se tiene  $T_1$  a  $T_d$  tuplas d-dimensionales que se etiquetan en  $c$  clases. Cada una de estas clases, tiene  $0$  a  $T$  tuplas. El vector  $\mathbf{S}$  se requiere, para encontrar  $\mathbf{U}^T \mathbf{L} \mathbf{S} = \mathbf{T}$  estas serán las proyecciones en baja dimensión de cada una de las tuplas.

Fisher-LDA en términos de optimización, pretende maximizar la función objetivo  $\mathbf{J}(\mathbf{S})$ , que se puede escribir de la siguiente manera:

$$3) \mathbf{J}(\mathbf{S}) = \frac{\mathbf{S}^T \mathbf{B} \mathbf{S}}{\mathbf{S}^T \mathbf{W} \mathbf{S}}; \mathbf{S} \in \mathbb{R}^d$$

$\mathbf{W}$  Es la matriz que garantiza la dispersión entre cada una de las clases,  $\mathbf{B}$  es la matriz que posee la dispersión inmersa en cada una de las clases a proyectar. También se puede notar de la siguiente forma:

$$4) \mathbf{W} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$$5) \mathbf{B} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

$\bar{\mathbf{x}}$  Es la media para cada una de las clases,  $\bar{\mathbf{x}}$  posee la media de todos los datos,  $N$  la cantidad de tuplas que pertenecen a la clase  $c$ . Fisher-LDA busca proyectar el vector  $\mathbf{w}$  que pueda maximizar el "cociente" entre la base de la matriz de dispersión inter-clase y la matriz de dispersión intra-clase.

Cuando  $S$  maximice la función objetivo tiene que satisfacer la siguiente condición:

$$6) \quad S^T L \tilde{A} S$$

Si  $S$  es no singular se puede desarrollar con las operaciones de valores propios para la matriz  $S^T S$ :

$$7) \quad S^T S L \tilde{A} S$$

Se sustituye el resultado para obtener el nuevo  $S$ :

$$8) \quad S = L \tilde{A} S$$

Del cual se tiene un  $S$  vector propio  $G$  de valor propio  $\lambda$ . Esto indica que para maximizar la función objetivo de la solución, se debe tener un vector propio asociado a mayor valor propio.

### III. MATERIALES & METODOS

Este trabajo analiza un conjunto de datos del repositorio UCI – Machine Learning Repository, esta matriz posee 100 registros y 13 atributos de naturaleza multivariante [12], se recomienda para tareas de clasificación por la naturaleza de los datos de tipo discreto. Esta base de datos cuenta con los resultados de un análisis químico de 100 vinos cultivados en la misma región en Italia, pero derivadas de tres cultivares diferentes [12]. El análisis determina las cantidades de 13 componentes que se encuentran en cada uno de los tres tipos de vinos, como se muestra en la siguiente tabla:

Data set Characteristics:	Multivariate
Attribute Characteristics:	Integer, Real
Associated Tasks:	Classification
Number of Instances:	178
Number of Attributes:	13
Missing Values?	No
Area:	Physical
Date Donated	1/07/1991
Number of Web Hits:	577701

Tabla 1. Caracterización de datos, Wine-DataSet

En la Tabla 1. Se tiene una matriz de datos donde se caracteriza los elementos del data-set, tales como atributos, tarea asociada, número de instancias, número de atributos, pérdida de datos, área de aplicación, fecha y número de visitas en la web. Los datos se representan con elementos en cada fila como una muestra (vino) y cada columna una variable medida. Notar que las variables medidas pueden ser de tipos distintos:

- **Instrumentales:** absorbancia en el infrarrojo, relación masa/carga en un espectro de masas...
- **Físicas:** densidad, índice de color...
- **Químicas:** concentración de cada compuesto o un índice global (por ejemplo acidez total o contenido en polifenoles)
- **Organolépticas:** tipos e intensidad de ciertos aromas o tipos e intensidad de sensaciones gustativas (generalmente se trata de medidas semicuantitativas basadas en escalas predefinidas)

De esta manera tenemos la matriz de datos representada en el espacio original o de alta dimensión de la siguiente manera:

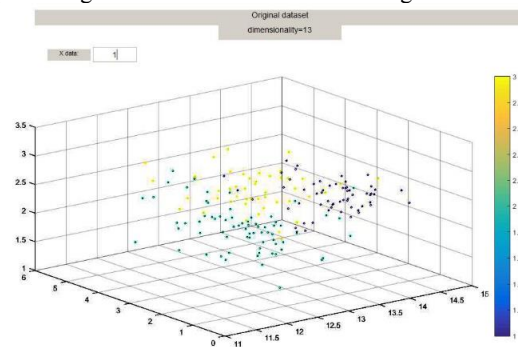


Figura 1. Matriz de datos de 13 dimensiones, con 100 registros, distribuidos en 3 clases que se muestran en colores diferentes.

En la Figura 1. Se observa cada vino se ubica en el gráfico como punto en un espacio de trece dimensiones. Es decir, su representación requiere trece ejes, el cual es imposible de visualizar. La tarea a seguir es, reducir la dimensión dada por los 13 atributos que componen el conjunto de datos y representarlos, en un espacio de menor dimensión para conservar sus propiedades y describir los elementos, con el mínimo de pérdida [13].

La solución para comprimir estos trece atributos a un espacio de menor dimensión es, caracterizando cada vino con menor número de atributos para conservar la información relevante de cada elemento, tal como se muestra en la figura 3. [13]

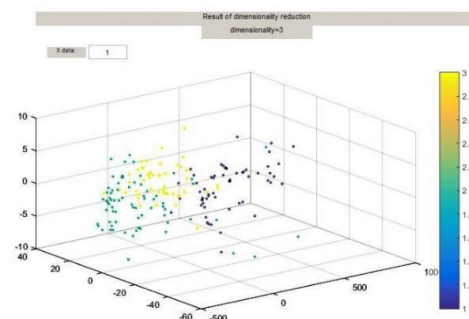


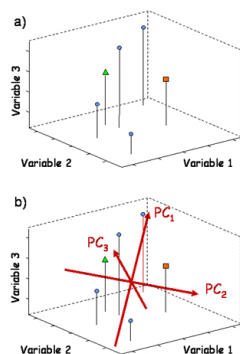
Figura 2. PCA reducción a espacio en 3 dimensiones.

En la Figura 2. se aprecia la reducción de la dimensión teniendo como base un espacio original de 13 dimensiones, a un espacio de 3 dimensiones que contiene la información que mejor describe los elementos del conjunto de datos, conservando la máxima varianza entre sus dos vectores propios.

MUESTRA	Var 1	Var 2	Var 3	Scores PC1	Scores PC2	Scores PC3
1	1.71	3.06	5.64	0.10	0.33	0.01
2	1.78	2.76	4.38	1.39	0.38	0.01
3	2.36	3.24	5.68	0.08	-0.33	-0.14
4	1.95	3.49	7.80	-2.07	-0.13	0.09
5	2.59	2.69	4.32	1.55	-0.41	0.08
6	1.76	3.39	6.75	-1.05	0.16	-0.05

**Tabla 2.** Valores de 3 variables o atributos para 6 vinos y scores de los 3 PC (Componentes Principales) calculados.

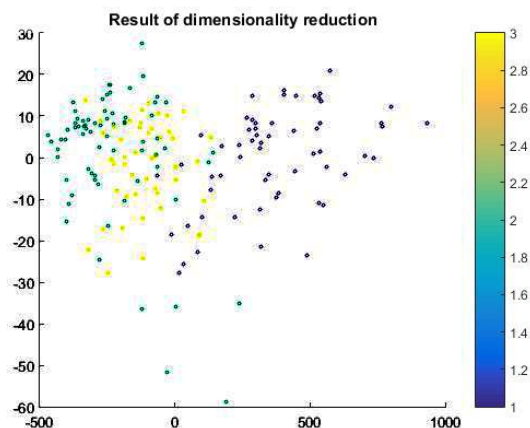
En la Tabla 2 muestra los valores de 3 atributos de 6 vinos de la Figura 2. Cada muestra está caracterizada por los valores medidos de estas variables. Se asignan cada variable a un eje de coordenadas obtenemos la representación de la Figura 3.



**Figura 3.**

a) Representación de los datos en el espacio tridimensional.

b) Los nuevos ejes (componentes principales, PC) son una rotación y traslación (los datos se centran previamente) de los ejes originales



**Figura 4.** PCA reducción a espacio en 2 dimensiones

En la Figura 4. se aprecia la reducción de la dimensión teniendo como base un espacio original de 13 dimensiones, a un espacio final de 2 dimensiones que contiene la información que mejor describe los elementos del conjunto de datos, conservando la máxima varianza entre sus dos vectores propios, del cual el primero vector recoge toda la información posible para describir la información y el segundo vector propio es ortogonal al anterior, recogiendo así la información faltante para describir 100% el conjunto de datos con su información.

En análisis discriminante lineal (LDA), se busca una variable  $U_L$  combinación lineal de las variables observadas  $U_L : \hat{O}$  que muestre las mayores diferencias entre las medias de los grupos de forma que nos permita la clasificación de uno de ellos con la máxima resolución posible [14] [13]. La media de los valores de la nueva variable para cada grupo es

$$9) \bar{U}_L = \frac{1}{n} \sum_{i=1}^n U_{Li} = \frac{1}{n} \sum_{i=1}^n \hat{O}_{Li}$$

Se trata, por tanto, hacer máximo  $\bar{U}_L$  sujeto a la restricción  $\bar{U}_L = 1$  para evitar las intermediaciones en los coeficientes producidos por la indeterminación en la escala de la variable combinada. La restricción significa que la variabilidad dentro de los grupos en la nueva variable es la unidad.

La solución viene dada por

$$10) \bar{U}_L = \frac{1}{n} \sum_{i=1}^n U_{Li} = \frac{1}{n} \sum_{i=1}^n \hat{O}_{Li}$$

Y la función discriminante lineal es:

$$11) U_L : \hat{O}_L = \frac{1}{n} \sum_{i=1}^n \hat{O}_{Li}$$

Se calcula calcular una combinación lineal de las 13 variables para este caso con la base de datos de vinos, para este proceso tomamos un coeficiente por cada variable o atributo y desarrollamos la sumatoria para todos los elementos posteriores. [15].

El proceso busca la transformación lineal que produce la máxima separación entre clases, para ello se calcula una nueva variable llamada discriminante, esta variable es la encargada de realizar el proceso de separación entre grupos. Posteriormente se calculan las medias de los diferentes grupos y la media del resultado entre estos, para obtener el punto de corte que será nuestra regla de decisión, que permite visualizar la clasificación de cada uno de los vinos en su respectivas poblaciones [16].

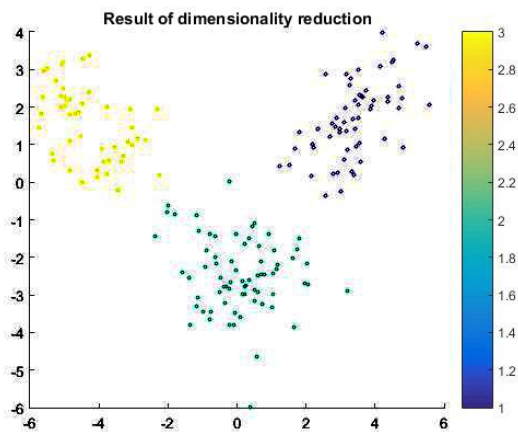


Figura 5. LDA reducción de 13 dimensiones a un espacio de 2 dimensiones

En la Figura 5. Se aprecia la reducción de la dimensión teniendo como base un espacio original de 13 dimensiones, a un espacio final de 2 dimensiones que contiene la información con la máxima separabilidad entre clases.

MUESTRAS PARA ESPACIO EN 2D	Var 1	Var 2	CombLine1	CombLine2
1	1.71	3.06	0.08	0.12
2	1.78	2.76	1.02	0.23
3	2.36	3.24	0.01	-0.13
4	1.95	3.49	-3.04	-0.19
5	2.59	2.69	0.98	-0.56
6	1.76	3.39	-0.02	1.05

Tabla 3. Valores de 2 variables o atributos para 2 Combinaciones lineales calculadas.

La Tabla 3 muestra los valores de 2 variables de 6 vinos de la Figura 5. Cada muestra está caracterizada por los valores medidos de estas variables.

La reducción con este método permite llevar un espacio de alta dimensión a un espacio de baja dimensionalidad, además de clasificar los datos linealmente separables. Se observa que hay pequeñas diferencias respecto a la tabla de valores de las matrices ya reducidas, del método Análisis de Componentes Principales, donde algunos valores muestra una tendencia a comprimir mejor los datos PCA, pero gráficamente no cumple con la tarea de separar las clases linealmente como LDA [10].

#### Comparación mediante Curvas de Calidad

Se analizan las curvas de calidad, del cual se toman las muestras de las matrices reducidas, utilizando una comparación de resultados del embejamiento representados con  $RNX(K)$ , que permite evaluar el área bajo la curva, tanto del lado derecho como del izquierdo, es decir cuando se asume que tiene una mejor representación en una topología global o local.

#### PCA VS LDA

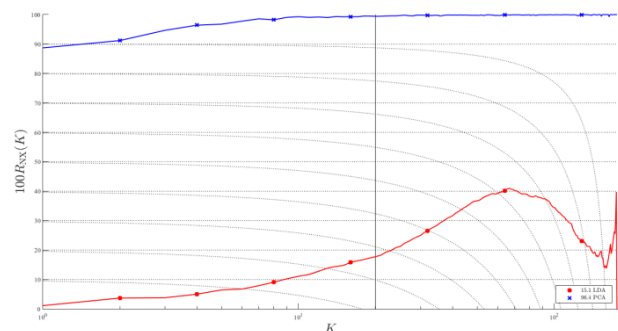


Figura 6. PCA versus LDA

●	15.1 LDA
×	96.4 PCA

La Figura 6. Muestra el método Análisis de Componentes Principales, del cual tiene una curva que se proyecta 9 unidades del eje y sobre Lineal Análisis Discriminante. Esto indica que la calidad de reducción de las 13 dimensiones con los 100 individuos y 3 poblaciones se comporta mejor para el método espectral PCA para compresión de datos, se aclara que este método no tuvo un buen rendimiento en separación de clases.

La segunda base de datos a evaluar es: cáncer de mama – breast cáncer.

Data set Characteristics:	Multivariate
Attribute Characteristics:	Categorical
Associated Tasks:	Classification
Number of Instances:	286
Number of Attributes:	9
Missing Values?	Yes
Area:	Life
Date Donated	11/07/1988
Number of Web Hits:	196504

Tabla 4. De caracterización de datos, Breast Cancer-DataSet

La base de datos, cáncer de mama posee 9 dimensiones representadas en atributos, la característica de estos atributos es categórico multivariado, el número de instancias o individuos que presenta este dataset es 286, con pérdida de valores confirmada según la tabla de caracterización de la base de datos. Los criterios de análisis para estos datos son curvas de calidad, tomando muestras de las matrices reducidas, para realizar una comparación de resultados del embejamiento representados con  $RNX(K)$ , que permite evaluar el área bajo la curva, del cual asume una mejor representación en una topología global o local.



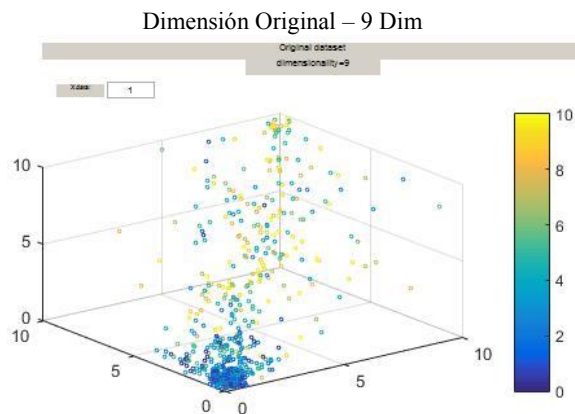


Figura 7. Breast Cancer – DataSet en alta dimensión.

La Figura 7. Muestra los datos representado del espacio original que posee 9 dimensiones o atributos de un hiper plano, el cual se aplicara el proceso de reducción de la dimensión con los métodos LDA y PCA.

A continuación el desarrollo de PCA para el conjunto de datos. Reducción de la dimensionalidad con Análisis de componentes principales (PCA), de 9 dimensiones a una dimensión en 2D.

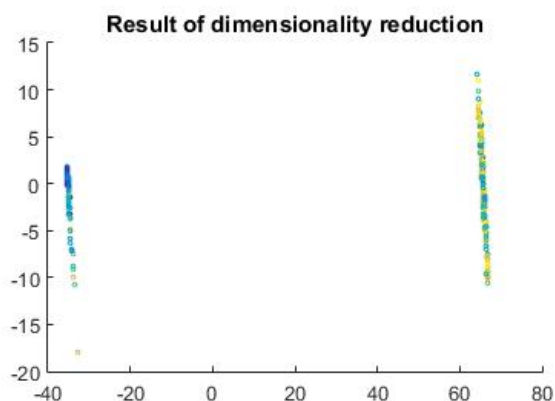


Figura 8. Breast Cancer – DataSet reducción de la dimensionalidad de 9 dimensiones a 2 dimensiones.

En la Figura 8. Se presenta la reducción de la dimensionalidad con Análisis discriminante lineal (LDA), de 9 dimensiones a una dimensión en 2D.

Debido a la naturaleza de los datos se observa que la agrupación natural de las clases conforma la topología de dos grupos que están separados de tal modo, que invitan a pensar si existe una relación lineal dependiente tal cual se muestra en la distribución de las clases en la figura 8, esto por el patrón visual que conformaron las clases en el desarrollo de la reducción de la dimensión.

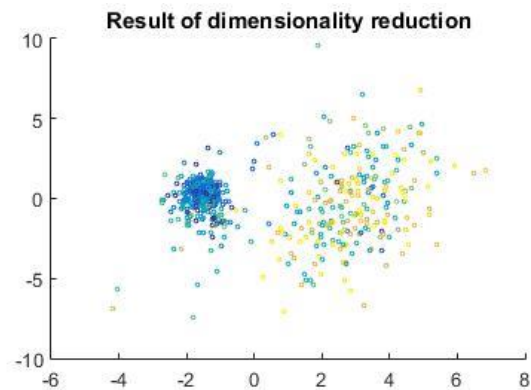


Figura 9. Breast Cancer – DataSet reducción de la dimensionalidad de 9 dimensiones a 2 dimensiones.

La Figura 9. Muestra los datos representado del espacio procesado a una dimensión menor que posee 2 dimensiones o atributos de un hiper plano, se puede concluir visualmente que los resultados para la tarea fundamental de este trabajo que es; “saber que método separa mejor las clases” es LDA.

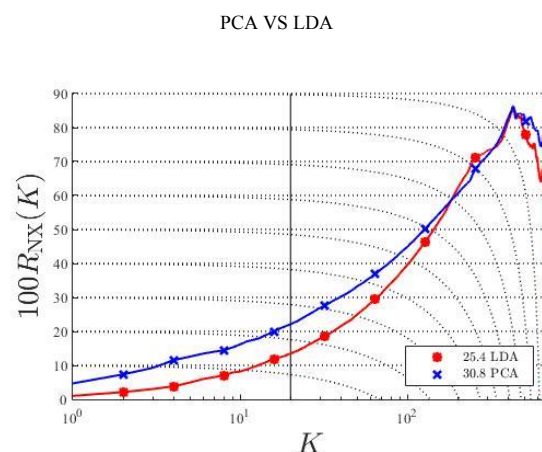


Figura 10. PCA versus LDA – Breast Cancer DataSet

En la Figura. 10 muestra que el método Análisis de Componentes Principales tiene una curva que se proyecta desde un valor cercano a Lineal Análisis Discriminante. Esto indica que la calidad de la reducción de las 9 dimensiones con los 569 individuos en 2 poblaciones se comporta mejor para el método PCA, pero visualmente el método LDA representa un mejor agrupamiento de los datos el cual permite claramente identificar los individuos pertenecientes a cada clase.

#### IV. CONCLUSIONES

Se concluye que Análisis Discriminante Lineal bajo el criterio de Fisher, dentro de su topología y modelo discriminante, obtiene resultados notoriamente efectivos en separabilidad intra clases, su fortaleza radica en proyectar un buen

desempeño en tareas de reducción de dimensión conservando la separabilidad entre clases como se aprecia en las gráficas de las figuras 6 y 10. LDA no posee un buen rendimiento en tareas de compresión como PCA [17], porque LDA desarrolla un proceso diferente el cual busca seguir patrones de distribución gaussianas para un buen funcionamiento, que garantice una transformación intra-clases para hallar la máxima dispersión entre cada clase y a su vez, reducir a su mínimo posible la dispersión interna de cada clase.

LDA tiene un mejor rendimiento en tareas de clasificación gracias al punto de corte del modelo discriminante que proyecta una robusta regla de decisión.

PCA posee un mejor desempeño que LDA en tareas de compresión de datos, sin embargo no es eficiente realizando tareas de separación entre clases [17], este método se caracteriza por realizar una combinación lineal, el cual proyecta los datos, sin perder todas sus propiedades para describirlos en un espacio de baja de dimensión.

## V. DISCUSIONES - TRABAJOS FUTUROS

Para trabajos futuros es necesario evaluar otros aspectos relevantes que conciernen a la temática de los algoritmos, ya que es necesario medir sus performance en complejidades en tiempo en ejecución, con datos masivos de fuentes, estructuradas y no estructuradas, en aras de establecer un acercamiento real con el entorno creciente de la IoT así como también en entornos BigData ya que a día de hoy son tendencias que promueven el ejercicio de la reducción de la dimensionalidad en pro de la eficiencia.

Dentro del contexto del estado de la técnica, es pertinente realizar una búsqueda de los nuevos métodos de reducción de la dimensión, en aras de dar información valiosa y relevante acerca de las tareas del mundo de hoy, ya que en datos masivos es una necesidad latente dar tratamiento especial para reducir el costo computacional así como también saber cuál será el más efectivo en tareas de alto performance en datos estructurados y no estructurados. El testing en complejidades de tiempo en ejecución sería un buen contraste para medir rendimiento y capacidad del algoritmo.

## VI. REFERENCIAS

- [1] P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation", *Neural computation*, vol. 15, April 2003.
- [2] J. Valencia-Aguirre, "Comparación de Métodos de Reducción de Dimensión Basados en Análisis por Localidades", *Tecno Lógicas* No. 25, May 2010.
- [3] Masashi Sugiyama, "Semi-supervised local Fisher discriminant analysis for dimensionality reduction", *Machine Learning is an international forum for research on computational approaches to learning* - Springer, June 2010.
- [4] M. Mauricio Delbracio, "Trabajo Final de Reconocimiento de Patrones Identificación utilizando PCA, ICA y LDA", Grupo de tratamiento de señales de la Universidad de la Republica - Insituto de Ingenieria Electrica, Montevideo - Uruguay, September 2006.
- [5] P. J. Husson François, "Exploratory Multivariate Analysis by Example Using R.", London, Chapman & Hall/CRC The R Series, p. 224p October 2009.
- [6] D. H. Peluffo-Ordóñez, "On the Relationship Between Dimensionality Reduction and Spectral Clustering from a Kernel Viewpoint", *Distributed Computing and Artificial Intelligence*, 13th International Conference - Springer, June 2016.
- [7] L. V. MAATEN, "Matlab Toolbox for Dimensionality Reduction", [En línea]. Available: <https://lvdmaaten.github.io/drttoolbox/>, January 2016.
- [8] MatLab, "MathWorks Software Theoretic FrameWork Dimensional Reduction", [En línea]. Available: <https://es.mathworks.com/products/matlab/>, August 2016.
- [9] D. H. Peluffo-Ordóñez, "Short review of dimensionality reduction methods based on stochastic neighbor embedding", *Advances in Self-Organizing Maps and Learning Vector Quantization*, November 2014.
- [10] G. M. Pimentel, "De la visualización a la sensorización de información", *Blucher Design Proceedings*, May 2013.
- [11] H. B. Ramires, "Descubrimiento de conocimiento en bases de datos: La perspectiva de la visualización inteligente de la información.", March 2014.
- [12] A. Inselberg, "Multidimensional Detective", *Information Visualization*, February 2007.
- [13] UCI - Repository of Machine Learning for scientific data sets, University Wisconsin, "<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>".
- [14] Y. Aliyari Ghassabeh, F. Rudzicz and H. A. Moghaddam, "LDA models", September 2010.
- [15] G. J. McLachlan, Fisher R.A, July 2004.
- [16] D. A. Keim, "Visual Database Exploration Techniques", *Exploring Databases*, vol. 7, March 2007.
- [17] W. S. Cleveland, "Visualizing Data", Horbart Press, June 1993.
- [18] A. I. and B. Dimsdale, "Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry", *IEEE Visualization*, October 2010.
- [19] A. Inselberg, "The Plane with Parallel Coordinates", *The Visual Computer*, April 1985.
- [20] P. E. Hoffman, "Table Visualizations: A Formal model and Its Applications", *University of Massachusetts*, February 2012.
- [21] H. P. Kriegel, "VisDB: Database Exploration Using Multimensional Visualization", *Computer Graphics and Applications*, March 1994.
- [22] C. A. Wistrand, "An Information Visualization and Exploration Environment", *Int'l Symp*, October 1995.
- [23] H. P. Kriegel, "Visualization Techniques for Mining Large Databases: A comparison", *IEEE Trans. Knowledge and Data Eng.*, pp. 923-936, April 1996.
- [24] D. Lalanne, "Surveying the complementary role of automatic data analysis and visualization in knowledge discovery", November 2013.
- [25] UCI Machine Learning Repository, breast cancer University Wisconsin "<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>".
- [26] J. Shlens. A, Principal Component Analysis a probabilistics models, arXiv preprint arXiv:1404.1100, April 2014.

[27] D. Kusnetzky, "What is Big Data?.Book Fundamentals Concepts ZDNet.", [En línea].Available:[Http://blogs.zdnet.com/virtualization/?p=178](http://blogs.zdnet.com/virtualization/?p=178), June 2013.

[28] P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation", Neural Computation, MIT Press Journals, pp. Pages 1373-1396, June 2006.