Pontificia Universidad Católica de Chile
Instituto de Ingenieria Matematica y Computacional
IMT2230 - Algebra lineal Avanzada y Modelamiento

# Proyecto N°2

9 de noviembre de 2023

Profesor Luis Cristobal Rojas - Ayudante Pablo Rademacher Barcelo

Rodrigo Andrés Martínez Becerra - Nicolás Alejandro Ortiz Muñoz - Daniel Alexis Cea Obando

---

## P1

**Respuesta 1.1:** Durante la realización de la Descomposición en Valores Singulares (SVD) de la matriz A en un espacio de dimensión k que mejor la aproxima, se evidencia cómo cada valor singular de A impacta en la precisión de la reconstrucción de cada imagen. Se destaca que, en el escenario en el que k adquiere el valor de 1, todas las imágenes representan la misma fotografía, dada la notable presencia de un valor singular que refleja la característica común que comparten, en este caso, la identificación de las imágenes como representaciones de mujeres. Por consiguiente, las proyecciones en el espacio de dimensión 1 exhiben notables similitudes entre sí. En contraste, al considerar valores de k iguales o superiores a 10, las imágenes manifiestan representaciones visualmente más fidedignas a sus respectivas formas originales.


Figura 1: SVD de dimensión k-esima

**Respuesta 1.2:** Después de llevar a cabo el Análisis de Componentes Principales (PCA) en la matriz A y proyectar las imágenes en el espacio de dimensión k que mejor las aproxima, se pueden discernir similitudes notables con la descomposición en Valores Singulares (SVD) en lo que respecta a cómo cada componente principal de A afecta la precisión de la reconstrucción de cada imagen. No obstante, se distinguen diferencias apreciables al realizar el PCA posterior a la centralización de las imágenes. Una de las observaciones más destacables radica en la apariencia de las imágenes reconstruidas. Al aplicar la centralización de las imágenes antes de llevar a cabo el PCA, se percibe que estas adquieren un matiz más opaco en contraste con las imágenes originales. Este cambio en la tonalidad es discernible en ejemplos específicos, como las imágenes 3 y 6 de la figura 2, aunque al revertir la centralización de las imágenes, este efecto no es evi dente.

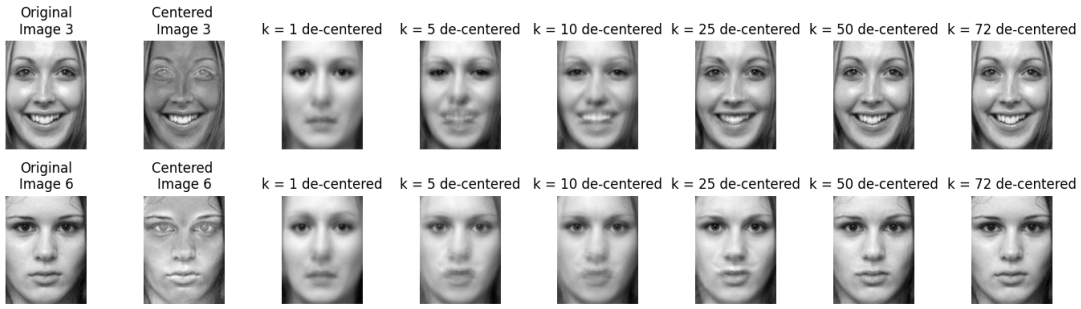


Figura 2: PCA de dimensión k

**Respuesta 1.3:** Al contrastar el Error Cuadrático Medio (ECM) para distintos valores de k, se evidencia que el Análisis de Componentes Principales (PCA) conlleva una leve mejora en la precisión en comparación con la Descomposición en Valores Singulares (SVD). No obstante, es pertinente resaltar que la centralización de datos, sea esta presente o no, no afecta de manera significativa el ECM resultante de PCA. En términos generales, se puede afirmar que PCA brinda una aproximación más precisa de los datos en cuanto al ECM en comparación con SVD.

| $k$ | ECM SVD | ECM PCA sin descentrar | ECM PCA descentrado | diff SVD y PCA |
|---|---|---|---|---|
| 1.0 | 1.1527 | 0.9759 | 0.9759 | 0.1768 |
| 5.0 | 0.6011 | 0.5721 | 0.5721 | 0.0290 |
| 10.0 | 0.3892 | 0.3797 | 0.3797 | 0.0094 |
| 25.0 | 0.1561 | 0.1511 | 0.1511 | 0.0050 |
| 50.0 | 0.0395 | 0.0374 | 0.0374 | 0.0022 |
| 72.0 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

Cuadro 1: Comparación de ECM entre SVD y PCA.

**Respuesta 1.4:**

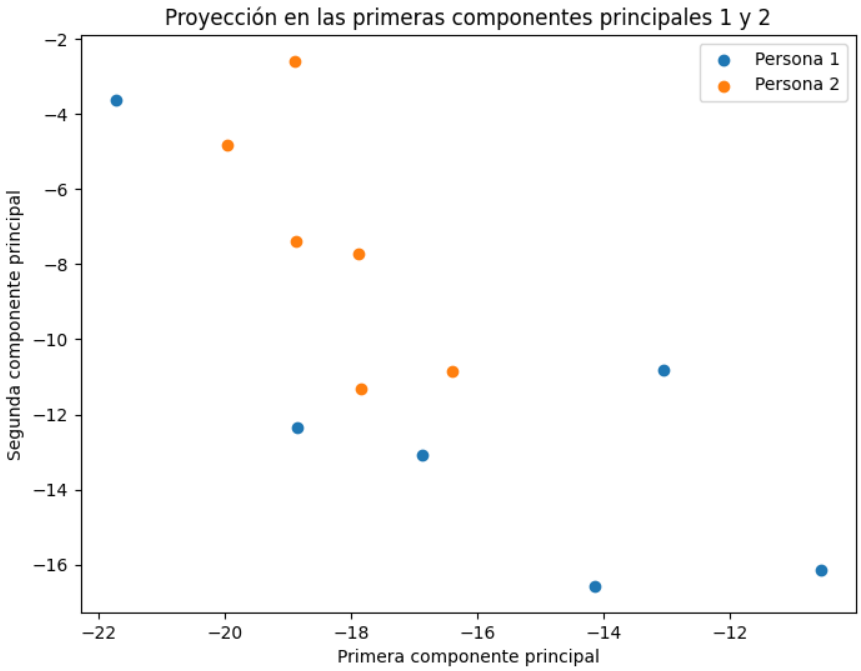


Figura 3: Grafico con las proyecciones de PCA1 y PCA2

En la representación gráfica previa, se aprecia una notable similitud en la proyección de las personas 1 y 2 en relación con las PCA1 y PCA2. Esta observación se sustenta en el hecho de que al considerar las primeras componentes de cada imagen (k=1 y k=2), se obtienen representaciones altamente semejantes entre sí, tal como se evidenció en las secciones 1.1 y 1.2. En estas secciones, se pudo constatar que al emplear un valor de k reducido, los rasgos faciales de las mujeres mantenían características similares.

**Respuesta 1.5:**

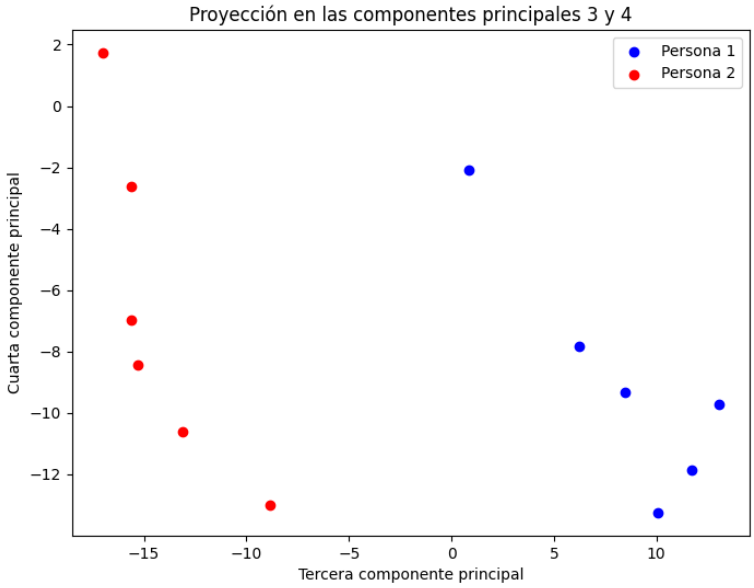


Figura 4: Grafico con las proyecciones de PCA1 y PCA2

Para distinguir entre las dos primeras personas, resulta más pertinente emplear el par de componentes 3 y 4, considerando que generalmente las dos primeras componentes capturan la mayor parte de la variabilidad de los datos. En contraste, las componentes 3 y 4 contienen una menor cantidad de información en comparación con las componentes 1 y 2, lo cual se refleja en cambios más sutiles en los rostros de las mujeres.

Este fenómeno se deriva de la observación de que PCA1 $\geq$ PCA2 $\geq$ PCA3 $\geq$ PCA4. Esta relación implica que al considerar un mayor número de componentes, se capturan detalles más sutiles que, no obstante, siguen permitiendo la distinción entre las personas.

**Respuesta 1.6:**

Para investigar la cuestión planteada en el punto 6, se procedió inicialmente a cargar la matriz test.npy para luego visualizar las seis imágenes asociadas. Posteriormente, se proyectaron estas imágenes sobre las 20 componentes principales iniciales obtenidas de la matriz A previamente mencionada, la cual comprende 72 rostros. La proyección resultante fue sometida a una evaluación exhaustiva con el propósito de analizar la calidad de las aproximaciones logradas. Es crucial analizar detalladamente las imágenes que exhiben una adecuada aproximación mediante el empleo de las componentes principales de A, junto con el discernimiento de los fundamentos subyacentes.

Específicamente, al examinar las aproximaciones de las imágenes proyectadas, se logró identificar aquellas que reflejan una marcada semejanza con sus contrapartes originales. Este fenómeno puede atribuirse a la habilidad de las primeras componentes principales para capturar de forma precisa los rasgos esenciales de los rostros presentes en el conjunto de datos. Por otro lado, aquellas imágenes que evidencian una discrepancia sustancial con sus contrapartes originales podrían indicar limitaciones intrínsecas en la capacidad de las componentes principales para representar de manera precisa ciertos objetos cotidianos incluidos en la matriz test.npy.

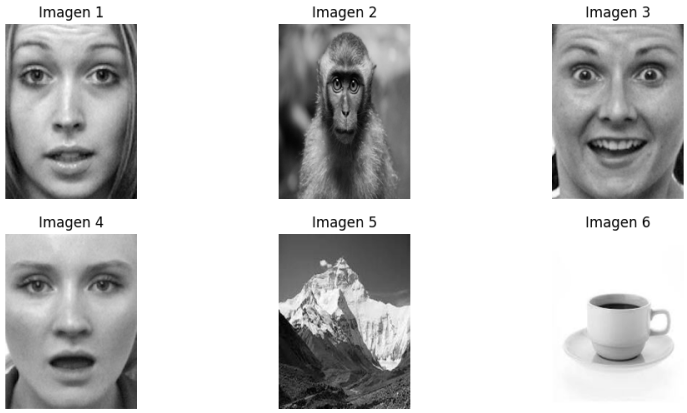


Figura 5: Imagenes del archivo test.npy

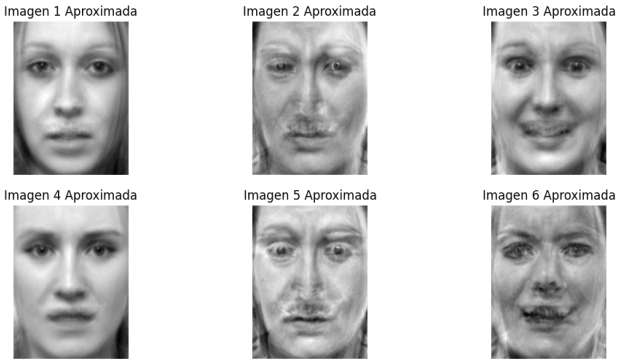


Figura 6: Proyecciones de las imagenes sobre las componentes de A

**Respuesta 1.7:** Se propone un enfoque integral para la clasificación precisa de imágenes faciales, el cual se fundamenta en la integración de redes neuronales convolucionales (CNN) y el análisis de componentes principales (PCA). Se reconoce el papel destacado de las CNN en la extracción meticulosa de atributos sofisticados de imágenes, lo que permite una clasificación robusta y precisa de rostros en el conjunto de datos. La arquitectura del modelo CNN se adapta estratégicamente a la complejidad de la tarea y a la naturaleza intrínseca de los datos de imágenes faciales, asegurando una representación efectiva de las características distintivas relevantes para la identificación de rostros.

El PCA, por otro lado, se plantea como una herramienta preliminar esencial para reducir la dimensionalidad de los datos, facilitando la proyección de los datos de imágenes en un espacio de menor dimensión que conserva la variabilidad crucial. Esto resulta beneficioso para la representación precisa de las características fundamentales de los rostros, permitiendo así la discriminación entre rostros y otros objetos cotidianos presentes en los conjuntos de datos.

La implementación práctica de este enfoque integral requiere una meticulosa compilación y etiquetado de conjuntos de datos de entrenamiento representativos, que abarquen la diversidad inherente de los rostros y objetos cotidianos. El modelo de CNN se somete a un riguroso proceso de entrenamiento y validación, utilizando conjuntos de datos independientes para garantizar su capacidad de generalización y precisión en la identificación de rostros en escenarios del mundo real.

En este contexto, se subraya la importancia crucial de consideraciones éticas y de privacidad al trabajar con datos de imágenes faciales, lo que implica una reflexión cuidadosa sobre los posibles sesgos y limitaciones tanto del conjunto de datos como del modelo propuesto. La implementación y evaluación de este enfoque exhaustivo requieren una atención detallada no solo a los aspectos técnicos, sino también a las implicaciones éticas y legales que aseguren la equidad y confiabilidad del sistema en el ámbito de la clasificación de imágenes faciales.

Además de las técnicas mencionadas anteriormente, es esencial considerar otros enfoques

de clasificación relevantes en la identificación de rostros, como el clasificador K-Nearest Neighbors (KNN), los árboles de decisión y las máquinas de vectores de soporte (SVMs). Estas técnicas proporcionan un marco amplio y completo para la identificación precisa y confiable de rostros en diversos contextos de ciencias de la computación y visión por computadora.

# P2:

**Respuesta 2.1:** Dado que en el contexto parlamentario, los votos se representan mediante los siguientes valores: 1 (apruebo), -1 (rechazo) y 0 (abstención), los cuales reflejan las decisiones de los diputados de ejercer su derecho al voto, en este contexto, el valor NaN se considera como un voto blanco, nulo o no participo, indicando que el diputado no ha emitido una decisión concreta en la votación o simplemente no estaba en la votacion. El analisis exploratorio es posible encontralor en el notebook adjunto.

**Respuesta 2.2a:** Dado que hemos optado por implementar la segunda de opcion de remplazo, los valores NaN seran sustituidos por el promedio de cada votacion; el siguiente algoritmo es el empleado para el remplazo:

```
def apply_replace(df, df_replace):
  df = df.copy()
  df_columns = df[df.columns[1:1034]]
  for index, row in df.iterrows():
    for col_name in df_columns:
      if pd.isna(row[col_name]):
        df.at[index, col_name] = df_replace.T[col_name]
  return df
```

Con el codigo anterior remplazamos los valores nulos y al resultado obtenido le aplicamos PCA, obteniendo el siguiente grafico:

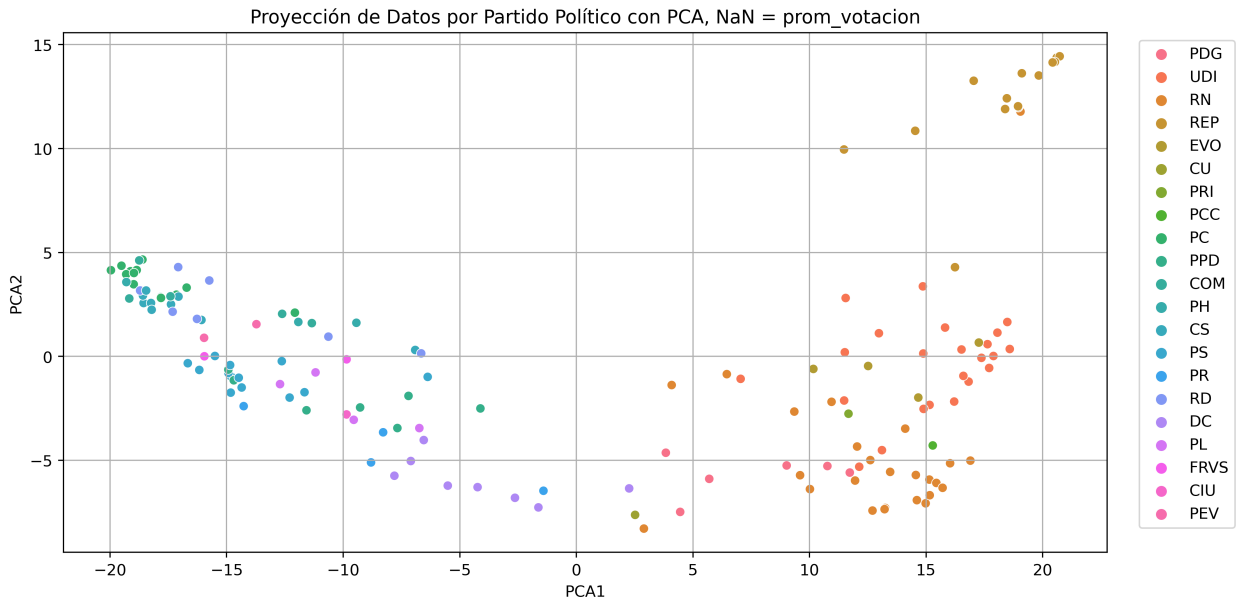


Figura 7: Grafico de PCA con valores NaN remplazados por 0

**Respuesta 2.2b:** Para reemplazar los valores NaN por el promedio de su partido político, se utilizará el siguiente algoritmo:

**Codigo:**

```
def apply_replace(df, df_replace):
  df = df.copy()
  df_columns = df.columns[1:1034]
  for index, row in df.iterrows():
    for col_name in df_columns:
      if pd.isna(row[col_name]):
        replace_col = df_replace[df_replace["party"] == row["party"]]
        df.loc[index, col_name] = replace_col.iloc[0][col_name]
  return df
```

Obteniendo un dataframe con los valores NaN reemplazados por el partido político del diputado/a que tenía valores nulos. Una vez agregados los datos, proyectamos los datos sobre la matriz de transformación obtenida al realizar la PCA. La grafica resultante del paso anterior es:

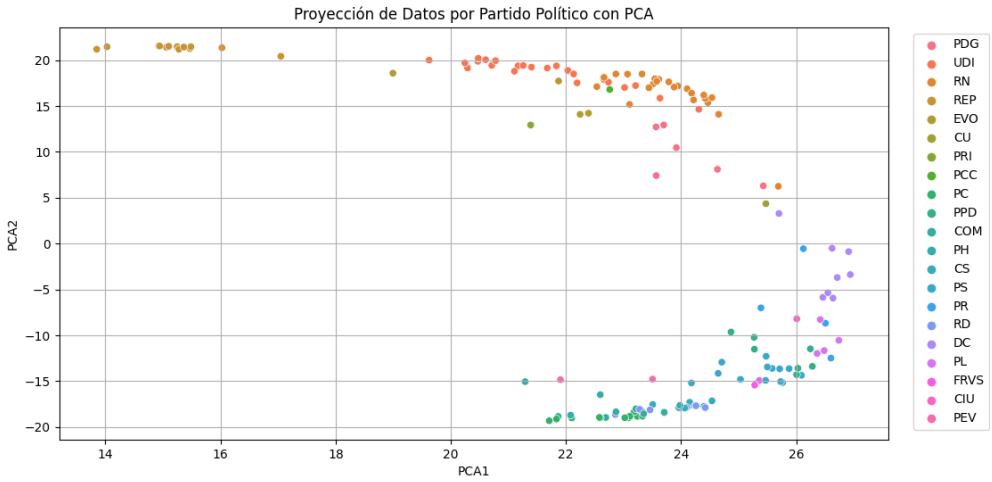


Figura 8: Grafico de PCA con valores NaN remplazados por el algoritmo; datos no centrados.

Sin embargo obtenemos un grafico que no es facilmente interpretable, por lo que las componentes principales seran multipicadas por -1. Ademas estos datos van a ser centrados usando '$X - X.mean()$'

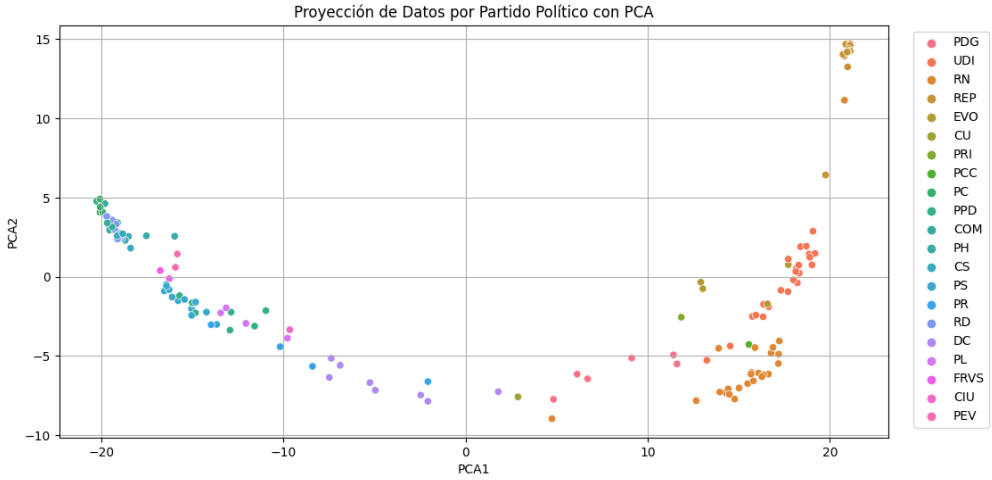


Figura 9: Grafico de PCA con valores NaN remplazados por el algoritmo; datos centrados.

**Respuesta 2.2c:** En ambas graficas se observa una separacion entre los partidos de izquierda y derecha. La derecha esta en el intervalo PCA1 $\in [0, 25]$ y la izquierda en PCA1 $\in [-25, 0]$, por lo que si existe una separacion entre los partidos politicos de Chile. En particular dicha separacion luce de la siguiente manera, y se expresa con PCA1.

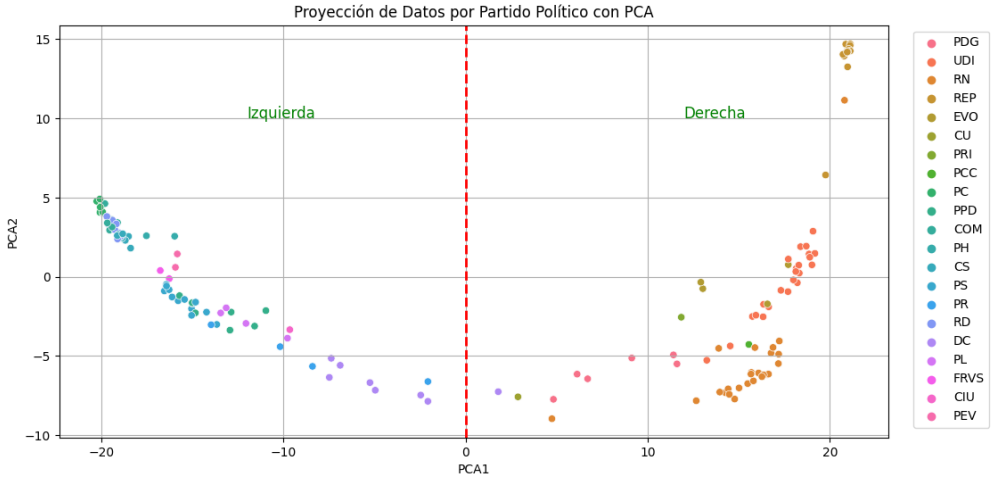


Figura 10: Grafica con la separacion por partidos politicos

El protocolo de reemplazo óptimo debe minimizar el sesgo en los datos. En este sentido, el tercer protocolo, que sustituye los valores nulos por el promedio del partido político al que pertenecen, es preferible al segundo protocolo, que solo utiliza el promedio de la votación en general. Esto se debe a que el tercer protocolo se acerca más a la decisión individual del diputado o diputada, teniendo en cuenta la afiliación política, en lugar de recurrir a valores promedios generales que podrían distorsionar la realidad de cada diputado/a.