Machine Learning y Data Analytics

HW 8 - Grupo 1

Ian Amighini
Julieta Brey
Lorenzo Nastri
Camila Sobrino
Matias Rodriguez Brun

Profesor Titular: Sergio Pernice

Consigna

1. Si corremos PCA a la matriz binaria X (o a X centrada). ¿Cuál sería la dimensión de los vectores outputs?

2. Examinaremos los primeros 2 componentes principales de X. Estos componentes contienen mucha información sobre nuestro conjunto de datos. Cree un diagrama de dispersión, con cada una de las 995 filas de X proyectadas en los primeros dos componentes principales. En otras palabras, el eje horizontal debe ser el espaneado por v1, el eje vertical por v2, y cada individuo debe proyectarse en el subespacio espaneado por v1 y v2. El gráfico debe usar un color diferente para cada población e incluir una leyenda.

3. Enumere 1 o 2 hechos básicos sobre el gráfico creado en la parte 2. ¿Puede interpretar los primeros dos componentes principales? ¿Qué aspectos de los datos capturan los dos primeros componentes principales? Sugerencia: piense en historia y geografía.

4. Ahora examinaremos el tercer componente principal de X. Cree otro diagrama de dispersión con cada individuo proyectado en el subespacio correspondiente a los componentes principales primero y tercero. Juegue con diferentes esquemas de etiquetado (con etiquetas derivadas de los metadatos) para explicar los grupos que ve. Su gráfico debe incluir una leyenda.

5. Algo debería llamar la atención en el gráfico del paso anterior. En una oración, ¿qué información captura el tercer componente principal?

6. En esta parte inspeccionará el tercer componente principal. Grafique el índice de nucleobase vs. el valor absoluto del tercer componente principal. ¿Nota algo? ¿Cuál sería una posible explicación? Sugerencia: piense en los cromosomas.

Proyecto Genoma

1_TP_PCA_ProyectoGenoma.ipynb

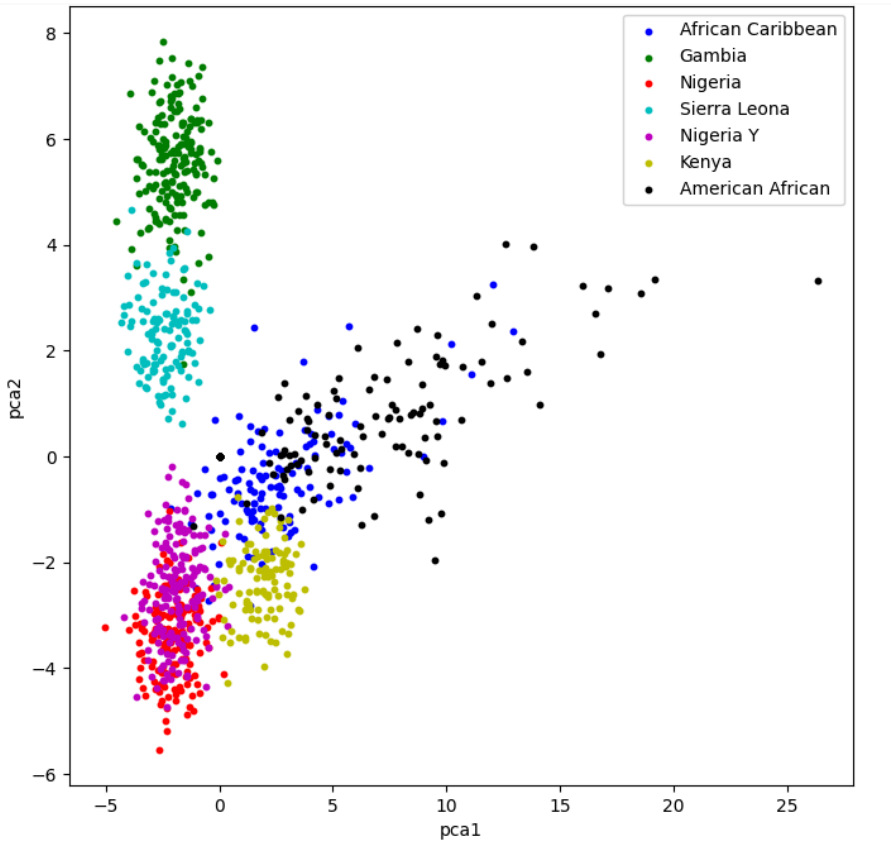
1. **¿Cuál sería la dimensión de los vectores outputs?**

```
data_dir = '/content/ML/'
data_filename = 'p4dataset2020.txt'

n_cols = 10101
n_rows = 995
n_metadata_cols = 3
```

Como X tiene forma (995×10101), el número de componentes principales (dimensión de cada "vector salida") es min(995, 10101) = 995. Dicho de otra forma, PCA sobre X genera 995 componentes (o, si se hace sobre la matriz de covarianza, 995 valores propios distintos de cero).

2. **Cree un diagrama de dispersión, con cada una de las 995 filas de X proyectadas en los primeros dos componentes principales.**



- Los primeros dos PCs capturan claramente la subdivisión geográfica dentro de África y distinguen poblaciones con historias demográficas y migratorias diferentes.

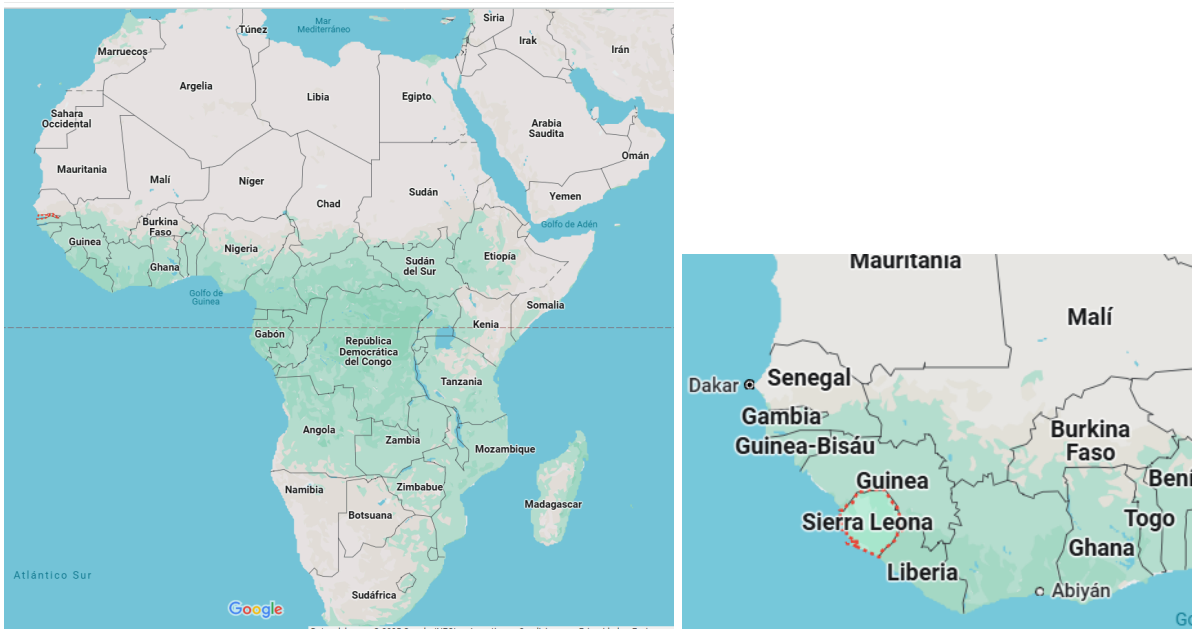
- Las poblaciones afro-descendientes en América y el Caribe aparecen intermedias y muy dispersas, lo que concuerda con la mezcla variable de linajes africanos y no-africanos en sus genomas.

- La PCA es una herramienta muy eficaz para visualizar patrones de variación genómica y detectar tanto estructuras poblacionales como grados de admixture.

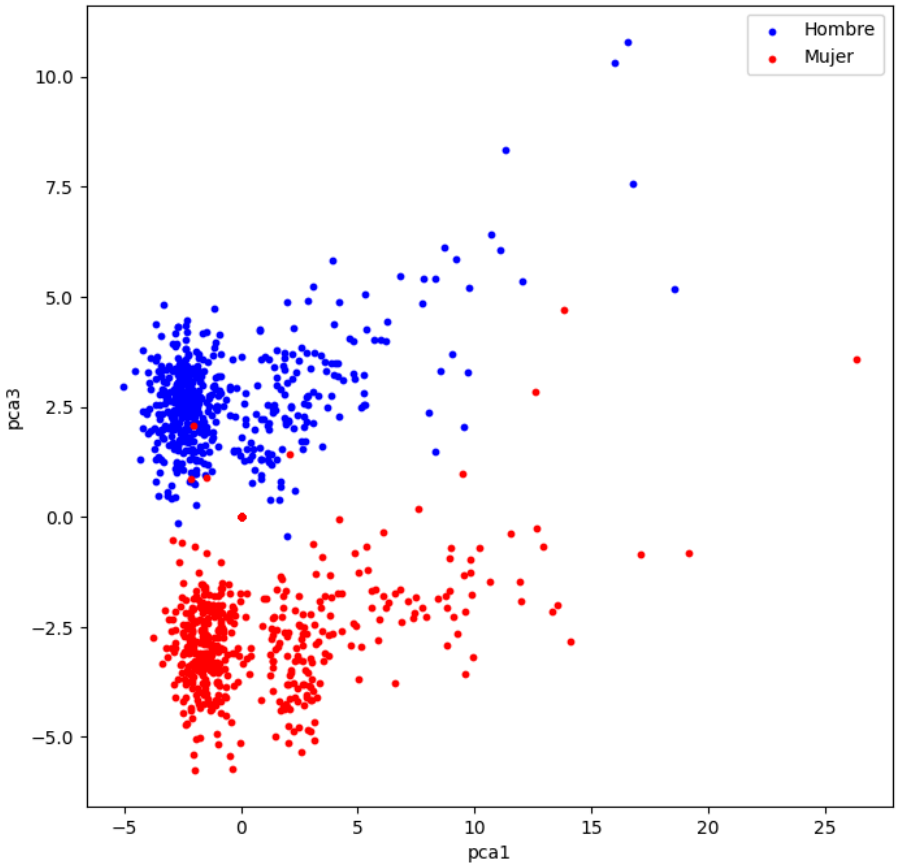
3. Interpretación de los componentes 1 y 2

- PC1: captura principalmente la separación África vs. no-África. Las coordenadas más negativas (o positivas, según orientación) de PC1 coinciden con todos los individuos de origen africano, mientras que los no africanos se sitúan al otro lado del eje. Esto refleja el fuerte gradiente genético histórico entre las poblaciones subsaharianas y aquellas de Europa/Asia, producto de divergencias evolutivas antiguas y barreras geográficas.

- PC2: discrimina variaciones dentro de África subsahariana. En el eje vertical (PC2), se distinguen sutilmente regiones como África occidental (Gambia, Nigeria) frente a África central-sur (Nigeria Y, Kenya). Esa “barra” vertical diferenciadora puede interpretarse como un efecto de migraciones locales y subdivisiones poblacionales históricas dentro del continente—por ejemplo, linajes occidentales vs. orientales—más allá de la separación continental que ya capta PC1.



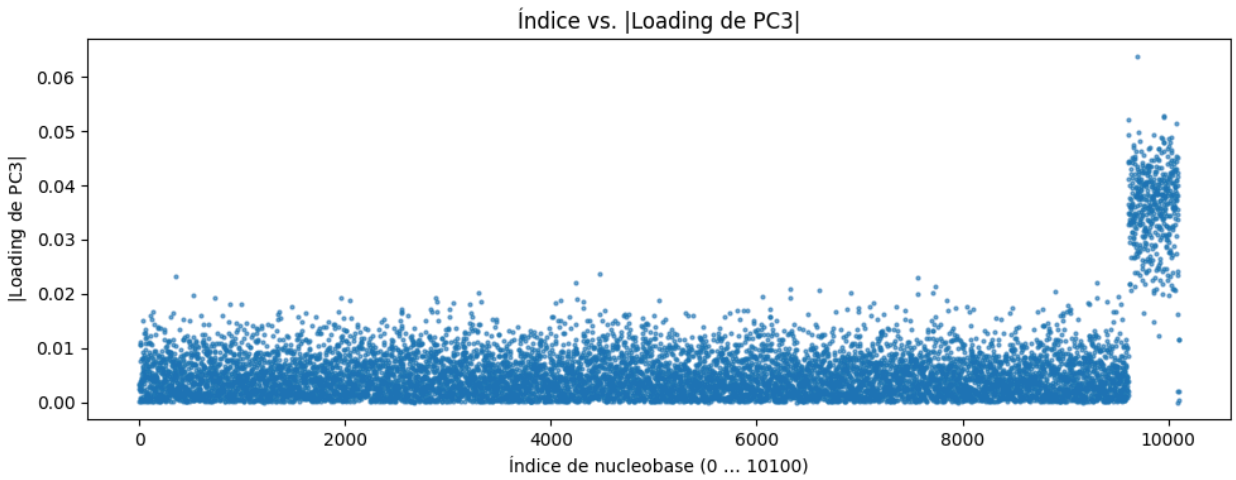
4. **Tercer componente principal de X. Cree otro diagrama de dispersión con cada individuo proyectado en el subespacio correspondiente a los componentes principales primero y tercero. Juegue con diferentes esquemas de etiquetado (con etiquetas derivadas de los metadatos) para explicar los grupos que ve. Su gráfico debe incluir una leyenda.**



5. **Algo debería llamar la atención en el gráfico del paso anterior. En una oración, ¿qué información captura el tercer componente principal?**

Los cromosomas sexuales (especialmente X) tienen un patrón de herencia diferente (los hombres tienen una y las mujeres dos copias), lo que produce una varianza distinta en los genotipos. Esa varianza “extra” hace que sus SNPs aparezcan con coeficientes grandes en PC3

6. **En esta parte inspeccionará el tercer componente principal. Grafique el índice de nucleobase vs. el valor absoluto del tercer componente principal. ¿Nota algo? ¿Cuál sería una posible explicación? Sugerencia: piense en los cromosomas.**



Al graficar el índice de cada nucleobase (0, 1, 2, …, 10 100) frente al valor absoluto de su coeficiente en el tercer componente principal (PC3), se observa lo siguiente:

a- Para los índices correspondientes a los cromosomas autosómicos (aprox. 0–9 400), los valores de |Loading de PC3| son muy bajos, cercanos a cero. Esto indica que, tras explicar la mayor parte de la variación por medio de PC1 y PC2, PC3 no capta prácticamente ninguna señal adicional en los marcadores autosómicos.

b- **A partir de un cierto índice (aprox. 9 400 en adelante, que coincide con el inicio de los SNPs del cromosoma X), los coeficientes absolutos de PC3 aumentan notablemente y forman un bloque continuo de valores altos**.

La explicación principal es que **PC3 está capturando la variación diferencial propia del cromosoma X, donde los hombres presentan una sola copia (ploidía 1) y las mujeres dos copias (ploidía 2).** Esta diferencia en variabilidad genética provoca que los SNPs del cromosoma X exhiban loadings mucho mayores en PC3 que los autosomas. En consecuencia, el segmento de índices correspondiente a las posiciones del cromosoma X aparece con un pico claro en el gráfico.

En resumen, el patrón observado confirma que *PC3 refleja principalmente la varianza asociada a los cromosomas sexuales, de modo que, si se desea limitar el análisis a la subestructura poblacional en autosomas, bastaría con omitir las columnas a partir del índice donde comienzan los SNPs de X*.