PROYECTO DE AGRUPAMIENTO

Tema: Implementación y evaluación de clustering con el algoritmo K-means en el dataset Adult (UCI Census Income)

1. Análisis Estadístico Descriptivo del Dataset Adult Census

1.1 Introducción y Contexto

El presente trabajo se enmarca en el análisis de datos socioeconómicos del dataset Adult Income, con el objetivo de explorar las posibles desigualdades laborales y de ingresos en función de características demográficas. En particular, se busca identificar patrones de agrupamiento entre individuos según su raza y género, considerando su ocupación y nivel de ingresos. Este enfoque permite examinar si ciertos grupos poblacionales presentan una mayor concentración en determinadas categorías ocupacionales o rangos salariales, lo cual podría reflejar diferencias estructurales en el mercado laboral. Para garantizar la validez poblacional de los resultados, se incorpora la variable fnlwgt únicamente en la fase interpretativa del análisis, utilizándola como ponderador muestral que refleja la representatividad real de cada registro dentro de la población total, sin intervenir en el proceso de agrupamiento ni en los cálculos descriptivos principales.

1.2 Preparación y Limpieza de Datos

Se consideraron todas las variables del dataset Adult Census, clasificándolas en:

Variables Numéricas (7): age, education_num, capital_gain, capital_loss, hours_per_week, income_binary, fnlwgt

Variables Categóricas (8): workclass, education, marital_status, occupation, relationship, race, sex, native_country

Decisión Metodológica: Se incluyo la variable `fnlwgt` (final weight) del análisis, ya que representa el peso muestral asignado a cada registro, indicando cuántas personas reales de la población están representadas por ese individuo en la muestra. Aunque no describe una característica personal, su presencia es necesaria porque permite ajustar los resultados del análisis a la estructura poblacional real. Sin esta variable, las proporciones y medias calculadas reflejarían solo la composición de la muestra, no la de la población de referencia. Por ello, fnlwgt se mantiene en el dataset para ponderar resultados en etapas de interpretación, asegurando que los análisis reflejen correctamente la representatividad poblacional, aun cuando no se use en el EDA ni en los modelos predictivos.

1.2.2 Análisis de Valores Faltantes

El análisis de integridad de los datos reveló la presencia de valores faltantes en tres variables categóricas, representados por el símbolo '?'. La evaluación sistemática de estos valores faltantes arrojó los siguientes resultados:

- **Caso 1 (3 columnas vacías):** 27 filas (0.08% del total)
- **Caso 2 (2 columnas vacías):** 1.809 filas (5.56% del total)

- **Caso 3 (1 columna vacía):** 563 filas (1.73% del total)

Distribución por variable:

- `workclass`: 1.836 valores faltantes (5.64%)
- `occupation`: 1.843 valores faltantes (5.66%)
- `native_country`: 583 valores faltantes (1.79%)

Estrategia de Limpieza Implementada: Dado que el porcentaje total de valores faltantes supera el umbral crítico del 5% por variable individual, se implementó una estrategia de rellenado con valores modales para las variables categóricas:

- `workclass`: rellenado con 'Private' (moda)
- `occupation`: rellenado con 'Prof-specialty' (moda)
- `native_country`: rellenado con 'United-States' (moda)

1.3 Análisis Estadístico Descriptivo

1.3.1 Medidas de Tendencia Central

El análisis de las medidas de tendencia central permite describir el comportamiento promedio de las principales variables numéricas del conjunto de datos.

La edad promedio de los individuos es de 38,58 años, con una mediana de 37 años, lo que sugiere una distribución relativamente simétrica con predominio de adultos en edad laboral activa.

En cuanto a la variable education_num, su valor promedio (10,08) y mediana (10) indican que la mayoría de los individuos alcanzó un nivel educativo correspondiente a la categoría "Some-college", según la codificación del dataset Adult Census. Esto representa personas que han completado la educación secundaria y cursado parcialmente estudios terciarios o universitarios, sin haber obtenido un título formal.

Las ganancias y pérdidas de capital presentan medias de 1.077,65 y 87,30 dólares respectivamente, mientras que sus medianas son cero, evidenciando una fuerte asimetría positiva: la mayoría de los individuos no registra operaciones de capital, aunque un pequeño grupo concentra valores elevados.

El promedio de horas trabajadas por semana (40,44) y su mediana (40) indican una jornada laboral típica a tiempo completo.

En cuanto a la variable fnlwgt, su media (189.778,37) y mediana (178.356,00) indican la distribución de los pesos muestrales asignados a cada observación, con una moda de 123.011, reflejando la representatividad poblacional de cada registro en la muestra.

Finalmente, el 24,08% de los individuos reporta ingresos superiores a 50.000 dólares anuales, lo cual evidencia una marcada desigualdad en la distribución del ingreso.

1.3.2 Medidas de Dispersión

Las medidas de dispersión evidencian la variabilidad de las observaciones. La edad muestra un rango amplio (17 a 90 años) y una desviación estándar de 13,64, indicando una diversidad etaria considerable.

Los años de educación presentan menor dispersión (DE = 2,57), reflejando cierta homogeneidad educativa dentro del conjunto.

Las ganancias y pérdidas de capital exhiben alta dispersión (DE = 7.385,29 y 402,96 respectivamente), lo que refuerza la presencia de valores extremos asociados a individuos con inversiones o pérdidas significativas.

Por su parte, las horas trabajadas por semana muestran una desviación estándar de 12,35, que evidencia variabilidad en los regímenes laborales (empleos de tiempo parcial, completo o con horas extraordinarias).

La variable fnlwgt presenta una desviación estándar de 105.549,98, con un coeficiente de variación del 55,62%, indicando una alta variabilidad en los pesos muestrales que refleja la diversidad en la representatividad poblacional de las observaciones.

1.3.3 Análisis de Outliers

El análisis mediante boxplots permitió identificar valores atípicos en distintas magnitudes.

Las variables capital_gain (8,33%), capital_loss (4,67%) y especialmente hours_per_week (27,66%) concentran una proporción considerable de outliers.

La variable hours_per_week presenta la mayor proporción de valores atípicos, lo que evidencia una alta variabilidad en las horas trabajadas. Este comportamiento puede atribuirse a la existencia de diferentes modalidades laborales, tales como empleos de tiempo parcial, tiempo completo y con horas extraordinarias, reflejando la diversidad en las condiciones de empleo dentro de la muestra.

Por su parte, los outliers observados en capital_gain y capital_loss corresponden a una minoría de individuos con movimientos de capital significativamente superiores al promedio.

Finalmente, la baja proporción de valores atípicos en age (0,44%) y education_num (3,68%) indica estabilidad en las dimensiones demográficas y educativas.

La variable fnlwgt presenta 992 outliers (3,05%), concentrados en el rango superior de los pesos muestrales, lo que indica la presencia de observaciones con mayor representatividad poblacional que el promedio.

1.4 Análisis de Variables Categóricas

1.4.1 Distribución Demográfica

Distribución por Sexo:

- Masculino: 66,92% (21.790 individuos)
- Femenino: 33,08% (10.771 individuos)

La muestra presenta una mayor proporción de hombres que de mujeres, lo cual podría incidir en la distribución de ingresos y ocupaciones observadas, especialmente en contextos donde existen diferencias estructurales en la inserción laboral y las oportunidades económicas por género.

Distribución por Ingresos:

- ≤50K: 75,92% (24.720 individuos)
- >50K: 24,08% (7.841 individuos)

La mayoría de los individuos (75,92%) percibe ingresos anuales iguales o inferiores a 50.000 dólares, mientras que solo una cuarta parte (24,08%) supera dicho umbral. Esta disparidad refleja una estructura económica desigual, donde predomina la población de ingresos bajos y medios, característica consistente con los datos censales de los Estados Unidos en 1994.

1.4.2 Características Socioeconómicas

Clase de Trabajo (workclass):

En términos ocupacionales, la mayoría de los individuos trabaja en el sector privado (75,34%), seguido por los autoempleados (7,80%) y empleados de gobiernos locales o estatales. Esto sugiere una estructura laboral concentrada en el ámbito privado con participación limitada del sector público.

Nivel Educativo (education):

Los resultados evidencian que los niveles educativos más frecuentes corresponden a HS-grad (32,25%) y Some-college (22,39%), lo cual refleja que una gran proporción de la población completó la educación secundaria o cursó parcialmente estudios superiores.

Los niveles universitarios y de posgrado (Bachelors, Masters, Doctorate) son minoritarios, aunque, como se analizará posteriormente, estos grupos presentan una mayor proporción de ingresos altos, confirmando la relevancia de la educación como factor socioeconómico.

Estado Civil (marital_status):

La estructura marital revela que el 46% de los individuos está casado con cónyuge presente, seguido por quienes nunca se casaron (32,81%) y los divorciados (13,64%). En conjunto, estas tres categorías agrupan más del 90% de la muestra, lo que sugiere una población predominantemente casada o soltera, con menor representación de personas separadas o viudas.

Ocupación (occupation):

La distribución ocupacional muestra una diversificación moderada del mercado laboral. Las categorías más frecuentes son profesiones especializadas (18,37%), trabajos técnicos o de reparación (12,59%) y puestos ejecutivos o administrativos (12,49%).

Esto sugiere una estructura laboral con predominio de ocupaciones calificadas y técnicas, aunque también se observa participación de sectores de servicios y tareas manuales.

Relación Familiar (relationship):

La estructura familiar se caracteriza por una mayor proporción de jefes de hogar masculinos (Husband, 40,51%), seguida por individuos que no pertenecen a un núcleo familiar (25,50%) y por hijos dentro del hogar (15,56%).

Estos datos reflejan una muestra compuesta principalmente por adultos casados y cabezas de familia, lo cual puede relacionarse con los patrones de ingresos y empleo observados.

Raza (race):

La composición racial del conjunto de datos está dominada por individuos de raza blanca (85,50%), seguidos por personas de raza negra (9,60%) y asiáticas o isleñas del Pacífico (2,85%). Esta distribución es coherente con la estructura demográfica de Estados Unidos en el período de referencia (1994).

País de Origen (native_country):

La mayoría de los individuos nació en los Estados Unidos (89,60%), mientras que el resto se distribuye entre diversos países de América Latina, Asia y Europa, como México (1,93%), Filipinas (0,93%), Alemania (0,50%) y Canadá (0,42%), entre otros.

Esta diversidad geográfica pone de manifiesto el carácter multicultural de la sociedad estadounidense, aunque con predominio de población nativa.

1.5 Análisis de Correlaciones

1.5.1 Matriz de Correlación General

La matriz de correlación de Pearson evidencia ausencia de correlaciones lineales entre las variables numéricas analizadas.

Según los criterios establecidos ($r \geq 0.8$ fuerte, $0.5 < r < 0.8$ débil, $r < 0.5$ sin correlación), todos los coeficientes se encuentran por debajo de 0.5.

1.5.2 Correlaciones por Grupo de Ingresos

Al segmentar la muestra según el nivel de ingresos, las correlaciones de Pearson continúan mostrando valores inferiores a 0.5, lo que indica ausencia de correlaciones lineales dentro de cada grupo.

En el grupo de individuos con ingresos >50K, la relación más destacada se presenta entre education_num y capital_gain ($r = 0.106$), lo que sugiere una asociación leve entre mayor nivel educativo y mayores ganancias de capital, aunque sin estar correlacionadas.

Asimismo, la relación negativa entre age y hours_per_week ($r = -0.127$) muestra una tendencia muy débil a que las personas de mayor edad trabajen menos horas semanales.

En el grupo de ingresos ≤50K, las relaciones son aún menores, reflejando comportamientos más heterogéneos y una estructura interna sin relaciones lineales detectables entre las variables numéricas.

1.6 Análisis de Contingencia

1.6.1 Relación Sexo-Ingresos

La relación entre sexo e ingresos evidencia una brecha de género significativa.

El 89,05% de las mujeres percibe ≤50K frente al 69,43% de los hombres, mientras que el 30,57% de los hombres supera los 50K, contra solo 10,95% de las mujeres.

Estos resultados confirman la existencia de desigualdades de género en los ingresos, posiblemente relacionadas con diferencias en ocupaciones, niveles educativos o tipos de jornada laboral.

1.6.2 Relación Educación-Ingresos

Se observa una asociación positiva entre nivel educativo e ingresos.

Las personas con Doctorate (74,09%), Prof-school (73,44%) o Masters (55,66%) presentan las mayores proporciones de ingresos >50K, mientras que los niveles educativos básicos como Preschool (0,16%) o 1st-4th (3,57%) muestran escasa participación en los niveles salariales altos.

Estos resultados refuerzan el papel de la educación como factor determinante en la movilidad económica y el acceso a mejores oportunidades laborales.

1.7 Análisis de Ponderaciones y Representatividad Poblacional

1.7.1 Metodología de Análisis Ponderado

Para garantizar la validez poblacional de los resultados, se implementó un análisis comparativo entre proporciones calculadas con y sin ponderación por `fnlwgt`. La ponderación permite ajustar los resultados para reflejar la representatividad real de cada grupo en la población, identificando sesgos de muestreo que pueden distorsionar las interpretaciones sobre las desigualdades estructurales.

1.7.2 Impacto de la Ponderación en las Desigualdades por Género

El análisis ponderado revela diferencias significativas entre las proporciones calculadas con y sin ponderación por `fnlwgt`:

Resultados por Sexo:

- **Mujeres**: 10.95% (no ponderado) → 10.82% (ponderado) = -1.11% de diferencia
- **Hombres**: 30.57% (no ponderado) → 30.10% (ponderado) = -1.57% de diferencia

Interpretación: Ambos géneros están ligeramente sub-representados en la muestra, con los hombres mostrando mayor sub-representación. La brecha de género en ingresos altos se mantiene prácticamente igual (19.62% vs 19.28%), indicando una desigualdad estructural consistente independientemente del método de cálculo.

1.7.3 Desigualdades Raciales y Representatividad Muestral

Los resultados por raza muestran patrones más complejos de representatividad:

Resultados (ordenados por magnitud de diferencia):

1. **Amer-Indian-Eskimo**: 11.58% → 12.44% = +7.44% (sobre-representados en muestra)
2. **Asian-Pac-Islander**: 26.56% → 27.62% = +3.97% (sobre-representados en muestra)
3. **Black**: 12.39% → 12.83% = +3.58% (sobre-representados en muestra)
4. **Other**: 9.23% → 9.07% = -1.63% (sub-representados en muestra)
5. **White**: 25.59% → 25.48% = -0.43% (sub-representados en muestra)

Interpretación: Los grupos raciales minoritarios (Amer-Indian-Eskimo, Asian-Pac-Islander, Black) están sobre-representados en la muestra, mientras que White y Other están sub-representados. Esto sugiere que la muestra incluye desproporcionadamente a miembros de minorías raciales con mayor probabilidad de tener ingresos altos.

1.7.4 Segregación Ocupacional y Patrones de Concentración

El análisis ponderado revela patrones de segregación ocupacional significativos:

Concentración Ocupacional por Género:

- **Mujeres**: Índice de concentración = 0.1590 (mayor concentración)
- **Hombres**: Índice de concentración = 0.1117 (mayor diversidad)

Patrones de Segregación:

- Las mujeres muestran mayor concentración en ocupaciones de servicios y administrativas
- Los hombres están más distribuidos pero dominan ocupaciones de manufactura, construcción y gerenciales
- Esta segregación horizontal contribuye a las diferencias salariales observadas

1.7.5 Análisis Trivariado: Intersección de Desigualdades

El análisis de la intersección entre género, raza e ingresos revela patrones complejos de desigualdad:

Combinaciones con Mayor Proporción de Ingresos Altos:

- **Hombres Asian-Pac-Islander**: 34.5% de ingresos altos
- **Hombres White**: 31.4% de ingresos altos

Combinaciones con Menor Proporción de Ingresos Altos:

- **Mujeres Other**: 5.1% de ingresos altos
- **Mujeres Black**: 5.6% de ingresos altos

Diferencia Máxima: 29.4 puntos porcentuales entre la combinación con mayor y menor proporción de ingresos altos.

1.7.6 Implicaciones Metodológicas y Estructurales

Sesgos de Muestreo Identificados:

- **Sobre-representación**: Grupos raciales minoritarios, niveles educativos bajos, ocupaciones militares
- **Sub-representación**: Población blanca, ocupaciones agrícolas y de manufactura, servicio doméstico privado

Impacto en la Interpretación:

- Las desigualdades estructurales persisten independientemente del método de cálculo
- La ponderación revela que ciertos grupos están desproporcionadamente representados en la muestra
- Los patrones de desigualdad observados pueden ser aún más pronunciados a nivel poblacional real

Validación de Patrones de Agrupamiento:

- Los análisis ponderados confirman la existencia de patrones sistemáticos de agrupamiento por género, raza y ocupación
- Las desigualdades identificadas reflejan verdaderas diferencias estructurales en el acceso a oportunidades económicas
- La segregación ocupacional horizontal contribuye significativamente a las brechas salariales observadas

1.7.7 Conclusiones del Análisis Ponderado

El análisis ponderado revela que las desigualdades estructurales en el mercado laboral estadounidense de 1994 persisten independientemente del método de cálculo, pero con magnitudes ajustadas que reflejan la realidad poblacional. La implementación de la ponderación no solo corrige sesgos de muestreo, sino que proporciona una base más sólida para identificar patrones de agrupamiento y concentración ocupacional que reflejan verdaderas diferencias estructurales en el acceso a oportunidades económicas.

Los hallazgos confirman la existencia de brechas significativas por género, raza y nivel educativo, mientras que la ponderación revela que ciertos grupos están desproporcionadamente representados en la muestra, sugiriendo que los patrones de desigualdad observados pueden ser aún más pronunciados a nivel poblacional real.

---

2. Implementación y Evaluación de la Distancia de Gower

2.1 Feature Engineering: Variable de Ahorro Neto

Previo al cálculo de la distancia de Gower, se implementó una estrategia de ingeniería de características mediante la creación de la variable `net_capital`, que combina las variables `capital_gain` y `capital_loss` en una única métrica de ahorro neto. Esta decisión metodológica permite representar el balance financiero total de cada individuo con una sola variable, reduciendo la dimensionalidad del problema y mejorando la interpretabilidad económica del análisis.

La distribución confirma que la mayoría de los individuos (87%) no registra movimientos de capital, mientras que una minoría presenta ganancias o pérdidas significativas.

2.2 Procesamiento de Variables según Tipo

Para el cálculo de la distancia de Gower se seleccionaron 12 variables que capturan los aspectos más relevantes para identificar patrones de desigualdad socioeconómica.

Variables Numéricas (4): Las variables `age`, `net_capital` y `hours_per_week` se normalizaron mediante el método min-max, transformando sus valores al rango [0,1]. Esto garantiza comparabilidad entre atributos de diferentes escalas y evita que variables con magnitudes elevadas (como `net_capital` con rango de 104.355 dólares) dominen el cálculo de disimilitud sobre variables con rangos más pequeños (como `age` con rango de 73 años).

Variable Ordinal (1): La variable `education` se codificó respetando el orden natural de los niveles educativos (Preschool=1, Doctorate=16),es por esto, que se decidió eliminar del estudio de Gower la

variable `education_num`. Luego se normalizó en el rango [0,1], asegurando una contribución proporcional al nivel académico alcanzado.

Variables Categóricas Nominales (6): Las variables `workclass`, `marital_status`, `occupation`, `relationship`, `race` y `native_country` se mantienen sin transformación. En la distancia de Gower se comparan por igualdad o diferencia: si dos observaciones comparten el mismo valor (por ejemplo, la misma raza), la distancia en esa variable es 0; si difieren, la distancia es 1.

Variables Binarias (2): La variable `sex` se codificó como binaria simétrica (Male=0, Female=1), mientras que `income` ya estaba codificada como binaria (≤50K=0, >50K=1). Ambas variables asignan igual peso a ambas categorías dentro del cálculo de distancias.

2.3 Implementación y Validación

Durante la validación con un subconjunto reducido de seis registros, la matriz de distancias obtenida cumplió con las propiedades esperadas: simetría, diagonal nula y valores dentro del rango [0,1]. Los resultados reflejaron coherencia semántica en el cálculo de distancias.

Los valores de distancia muestran coherencia semántica: individuos con características similares presentan distancias bajas (0.2919), mientras que perfiles muy diferentes exhiben distancias altas (0.8233). La distribución evidencia variabilidad adecuada para identificar patrones de similitud y diferencia en el conjunto de datos.

2.4 Análisis de Matrices de Distancias

Al calcular la matriz completa para una muestra de 200 observaciones, se obtuvo una distribución de distancias con media 0.3817 y desviación estándar 0.1253, concentrada mayormente entre 0.25 y 0.50. Esto indica que la mayoría de los individuos presenta diferencias moderadas entre sí, con una dispersión que refleja la diversidad sociodemográfica del conjunto.

Análisis de Casos Extremos: El par más similar (distancia: 0.0011) corresponde a dos hombres blancos con ocupación Craft-repair e ingresos >50K, evidenciando que la distancia de Gower identifica correctamente individuos con perfiles socioeconómicos casi idénticos. Por su parte, el par más diferente (distancia: 0.7987) contrasta una mujer blanca en Prof-specialty con ingresos >50K contra un hombre asiático-pacífico en Adm-clerical con ingresos ≤50K, reflejando diferencias sustanciales en género, raza, ocupación e ingresos.

2.5 Análisis Exploratorio de Patrones por Grupos

El análisis de distancias promedio dentro y entre grupos revela patrones sistemáticos de segregación y homogeneidad socioeconómica que reflejan diferencias estructurales en el acceso a oportunidades económicas.

Distancias por Género:

La distancia promedio entre géneros es aproximadamente 38% mayor que las distancias promedio dentro de cada género, confirmando que hombres y mujeres exhiben perfiles socioeconómicos distintivos. Esta diferencia sugiere patrones de segregación ocupacional y diferencias en los niveles de ingresos y oportunidades laborales entre géneros.

Distancias por Nivel de Ingresos:

Los individuos con ingresos >50K presentan mayor homogeneidad interna (distancia promedio menor), lo que sugiere un perfil socioeconómico más consolidado y similar entre individuos de este grupo. La marcada diferencia entre niveles (distancias entre niveles 63% mayores que dentro de niveles) confirma la estructura de estratificación económica presente en el dataset.

Distancias por Ocupación (Top 5):

Las ocupaciones técnicas y de reparación muestran mayor cohesión interna, posiblemente asociada a perfiles demográficos y educativos similares entre quienes ejercen estas ocupaciones. Por el contrario, las ocupaciones administrativas presentan mayor diversidad en sus características, reflejando una composición más heterogénea del grupo.

3. Análisis de Escalabilidad Computacional

3.1 Diseño Experimental

Para evaluar la escalabilidad de la implementación de la distancia de Gower, se seleccionaron dos muestras aleatorias del dataset Adult Census: una de 5.000 registros y otra de 10.000 registros. Ambas muestras fueron procesadas con el mismo tratamiento de variables aplicado en el ejercicio 2, utilizando la misma semilla aleatoria (`random_state=42`) para garantizar reproducibilidad.

3.2 Resultados de Escalabilidad Computacional

Características Particulares:

La implementación muestra un consumo prácticamente constante de memoria por distancia calculada:

- Muestra 5K: 15,26 MB por millón de distancias
- Muestra 10K: 15,27 MB por millón de distancias

La estabilidad de esta métrica indica que la implementación maneja eficientemente el crecimiento de los datos sin pérdidas adicionales de memoria por overhead.

3.3 Limitaciones y Trade-offs Computacionales

Consideraciones Prácticas:

El análisis de escalabilidad revela que, si bien la implementación es eficiente, el crecimiento cuadrático impone limitaciones prácticas claras:

- Una muestra de 5K registros requiere aproximadamente 25 minutos de procesamiento y 191 MB de memoria
- Una muestra de 10K registros requiere aproximadamente 115 minutos de procesamiento y 763 MB de memoria
- La extrapolación a muestras mayores (20K, 50K) resultaría en tiempos y recursos computacionales prohibitivos con la implementación actual

Estos resultados establecen límites operativos razonables para el uso de distancia de Gower en aplicaciones prácticas. Por encima de 20K registros, se vuelve necesario recurrir a técnicas de escalado,

muestreo estratégico o implementaciones optimizadas para mantener tiempos y recursos aceptables.

4. Implementación y Evaluación del Algoritmo K-means

4.1 Metodología y Configuración Experimental

Para la aplicación del algoritmo K-means se seleccionaron cinco variables numéricas del dataset Adult Census: `age`, `education_num`, `capital_gain`, `capital_loss` y `hours_per_week`. Dado que K-means utiliza la distancia euclidiana como medida de similitud, las variables deben estar en una escala comparable para evitar que aquellas con valores más grandes dominen el proceso de agrupamiento. Por ello, se implementó una estrategia de normalización diferenciada, adaptada a la distribución de cada variable.

Las variables `age` y `hours_per_week` presentan distribuciones relativamente simétricas y con baja presencia de valores extremos. En estos casos se aplicó StandardScaler, una técnica que centra los datos en torno a su media y los escala según la desviación estándar.

Por otro lado, las variables `education_num`, `capital_gain` y `capital_loss` mostraron distribuciones marcadamente sesgadas y con valores atípicos significativos. Para ellas se utilizó MinMaxScaler, que transforma los datos al rango [0,1]. En particular, `capital_gain` y `capital_loss` contienen una gran cantidad de ceros y pocos valores muy altos, por lo que su normalización mediante MinMax resultó esencial para mantener la estabilidad del algoritmo.

Adicionalmente, se realizó feature engineering mediante la creación de la variable `capital_neto`, que combina `capital_gain` y `capital_loss` en una única métrica de balance financiero (`capital_neto = capital_gain - capital_loss`). Esta transformación reduce la dimensionalidad del problema y mejora la interpretabilidad económica del análisis, ya que representa el balance financiero total de cada individuo.

4.2 Determinación del Número Óptimo de Clusters

Los resultados mostraron una convergencia notable entre las métricas de Silhouette y Davies-Bouldin.

La coincidencia entre ambas métricas y la estabilidad entre muestras de diferente tamaño confirma la robustez de la solución k=5 como número óptimo de clusters.

4.3 Análisis de Patrones en los Clusters

Caracterización de los Clusters Identificados:

El análisis de los cinco clusters óptimos reveló patrones socioeconómicos distintivos:

Cluster 0 - Perfil de Adultos con Ingresos Moderados:

- Edad promedio: 49.6 años
- Educación promedio: 10.2 años
- Horas/semana: 41.8 (jornada completa)
- Capital neto promedio: 1.515.1 dólares
- % Income >50K: 39.3%
- Interpretación: Representa adultos en plena actividad laboral con ingresos medios-altos

Cluster 1 - Perfil de Jóvenes con Ingresos Muy Bajos:

- Edad promedio: 23.8 años
- Educación promedio: 9.6 años
- Horas/semana: 21.3 (empleos de tiempo parcial)
- Capital neto promedio: 182.5 dólares
- % Income >50K: 3.2%
- Interpretación: Jóvenes en etapas tempranas de carrera laboral con ingresos mínimos

Cluster 2 - Perfil de Adultos Mayores con Ingresos Bajos:

- Edad promedio: 62.9 años
- Educación promedio: 9.4 años
- Horas/semana: 18.3 (jornada reducida)
- Capital neto promedio: 576.7 dólares
- % Income >50K: 14.0%
- Interpretación: Adultos mayores con menor actividad laboral y bajos ingresos

Cluster 3 - Perfil de Trabajadores Intensivos con Ingresos Altos:

- Edad promedio: 39.4 años
- Educación promedio: 11.0 años
- Horas/semana: 63.8 (jornada extendida)
- Capital neto promedio: 3.108.3 dólares
- % Income >50K: 43.9%
- Interpretación: Profesionales con alta dedicación laboral y mejores ingresos

Cluster 4 - Perfil de Adultos Jóvenes con Ingresos Moderados:

- Edad promedio: 29.8 años
- Educación promedio: 10.0 años
- Horas/semana: 41.4 (jornada completa)
- Capital neto promedio: 575.9 dólares
- % Income >50K: 16.0%
- Interpretación: Adultos jóvenes en desarrollo profesional con ingresos medios

4.4 Limitaciones del Enfoque K-means

Pérdida de Información por Variables Categóricas

El análisis exclusivo de variables numéricas conlleva limitaciones significativas. Las variables excluidas incluyen información fundamental sobre características socioeconómicas: `workclass`, `education`, `marital_status`, `occupation`, `relationship`, `race`, `sex` y `native_country`.

Impacto en la Calidad del Clustering:

Esta exclusión resulta en una pérdida aproximada del 60% de la información disponible en el dataset, lo que se traduce en clusters menos interpretables desde una perspectiva socioeconómica. La separación obtenida es menos precisa entre grupos demográficos y se pierden patrones importantes relacionados con género, raza y ocupación.

5. Clustering Jerárquico Aglomerativo con Distancia de Gower

Nota metodológica: El análisis principal de este ejercicio se realizó mediante el script `scriptFinal.ipynb`. En la consigna del proyecto se solicitaba inicialmente el uso de linkage='average'. El script `script1.ipynb` fue desarrollado siguiendo esta especificación original y demostró que este parámetro producía clusters extremadamente desbalanceados (ratio >4999x), con distribuciones del tipo 99.98% vs 0.02%, lo que hacía imposible la interpretación de desigualdades demográficas. Por esta razón, se desarrolló `scriptFinal.ipynb` utilizando linkage='complete', que produjo clusters más balanceados y sustentables para el análisis socioeconómico, justificando así la decisión metodológica final adoptada.

5.2 Resultados del Clustering

El clustering jerárquico aglomerativo se aplicó sobre las matrices de distancia de Gower pre-calculadas en el ejercicio 3, utilizando las mismas muestras de 5.000 y 10.000 registros para garantizar consistencia metodológica.

Muestra de 5.000 registros:

- Silhouette Score: 0.2291
- Davies-Bouldin Index: 2.0366
- Distribución de clusters:
  - Cluster 0: 2.497 registros (49.94%)
  - Cluster 1: 116 registros (2.32%)
  - Cluster 2: 447 registros (8.94%)
  - Cluster 3: 1.940 registros (38.80%)
- Ratio de desbalance: 21.53x

Muestra de 10.000 registros:

- Silhouette Score: 0.2437
- Davies-Bouldin Index: 1.7959
- Distribución de clusters:
  - Cluster 0: 4.746 registros (47.46%)
  - Cluster 1: 966 registros (9.66%)
  - Cluster 2: 3.980 registros (39.80%)
  - Cluster 3: 308 registros (3.08%)
- Ratio de desbalance: 15.41x

5.3 Caracterización de Clusters Identificados

El análisis descriptivo de los clusters reveló patrones socioeconómicos diferenciados:

Cluster 0 (49.9% en 5K, 47.5% en 10K) - Trabajadores de Bajos Ingresos:

- Compuesto mayoritariamente por mujeres (52%) y hombres (48%)
- Más del 80% de raza blanca
- Más del 99% con ingresos ≤50K
- Edad promedio: 34 años

- Perfil ocupacional: Profesionales especializados (18%), servicios varios (16%), administrativos (16%)

Cluster 1 (2.3% en 5K, 9.7% en 10K) - Mujeres Profesionales con Ingresos Mixtos:

- Predominantemente mujeres (94-99% según muestra)
- Más del 80% de raza blanca
- Ingresos >50K en 43% (5K) y 26% (10K)
- Edad promedio: 40-43 años
- Perfil ocupacional: Profesionales especializados (23-26%), ejecutivos/gerentes (16-19%), administrativos (16-20%)

Cluster 2 (8.9% en 5K, 39.8% en 10K) - Hombres Trabajadores Manuales:

- Exclusivamente hombres (100%)
- Más del 90% de raza blanca
- Ingresos equilibrados: 45% >50K (10K), 43% >50K (5K)
- Edad promedio: 40-44 años
- Perfil ocupacional: Reparación/artesanía (19%), profesionales especializados (18%), ejecutivos/gerentes (16%)

Cluster 3 (38.8% en 5K, 3.1% en 10K) - Grupos Minoritarios con Alto Ingreso:

- Compuesto por hombres (100% en 5K, 71% en 10K)
- Más del 90% de raza blanca
- Ingresos >50K en 100% (5K) y 88% (10K)
- Edad promedio: 43 años
- Perfil ocupacional: Profesionales especializados (46%), ventas (16%), reparación/artesanía (12%)

5.5 Análisis Ponderado con Final Weight (fnlwgt)

El análisis ponderado utilizando la variable `fnlwgt` mostró diferencias sutiles pero consistentes entre las distribuciones ponderadas y no ponderadas, afectando principalmente la composición por género y ocupación:

Impacto del Ponderador:

- Diferencias en distribución por sexo: hasta 2.1 puntos porcentuales
- Diferencias en ocupación: hasta 3.8 puntos porcentuales
- Diferencias en ingresos: menores a 0.3 puntos porcentuales

Estas variaciones confirman que el ponderador poblacional captura aspectos de representatividad geográfica y sociodemográfica que no están presentes en el análisis no ponderado, aunque el impacto en la estructura general de clusters es limitado.

6. Análisis Comparativo: K-means vs Clustering Jerárquico con Distancia de Gower

6.1 Configuración del Análisis Comparativo

Se implementó un análisis comparativo directo entre los dos enfoques de clustering evaluados en ejercicios anteriores: K-means sobre variables numéricas normalizadas y clustering jerárquico aglomerativo con distancia de Gower. Ambos métodos se aplicaron sobre las mismas muestras reproducibles de 5.000 y 10.000 registros, utilizando los parámetros óptimos identificados en cada caso: k=5 para K-means y k=4 con linkage='complete' para Gower+Agglomerative.

6.2 Comparación de Métricas de Calidad

K-means supera significativamente a Gower+Agglomerative en Silhouette Score (casi el doble en ambas muestras), mientras que Gower muestra un Davies-Bouldin Index superior en la muestra 10K pero inferior en la muestra 5K. En términos de eficiencia computacional, K-means es sustancialmente más rápido, ejecutándose 7x más rápido en la muestra 5K y 40x más rápido en la muestra 10K.

6.3 Análisis de Balance de Clusters

K-means produce clusters más balanceados que Gower+Agglomerative en ambas muestras. La distribución más equilibrada de K-means indica que este método produce agrupaciones más homogéneas en tamaño, mientras que Gower genera clusters con mayor variabilidad en su composición.

6.4 Separación de Ingresos por Cluster

El análisis de distribución de ingresos >50K reveló diferencias sustanciales en la capacidad de cada método para discriminar niveles socioeconómicos.

Gower+Agglomerative demuestra una capacidad significativamente superior para separar clusters por nivel de ingresos, con diferencias de casi 100 puntos porcentuales entre el cluster de menor y mayor ingreso. Esta capacidad discriminativa se atribuye a la incorporación de variables categóricas (ocupación, educación, género, raza) que K-means no considera.

Conclusiones Generales

Síntesis de Hallazgos sobre Desigualdades Demográficas

El presente trabajo ha permitido identificar patrones sistemáticos de desigualdad en el mercado laboral estadounidense de 1994 que trascienden las características individuales y reflejan diferencias estructurales en el acceso a oportunidades económicas. Los análisis realizados confirman la existencia de múltiples dimensiones de desigualdad que operan de manera interconectada.

Desigualdades por Género: La brecha salarial identificada es pronunciada y persistente: mientras que el 30.57% de los hombres supera los 50.000 dólares anuales, solo el 10.95% de las mujeres alcanza este umbral. Esta diferencia de casi 20 puntos porcentuales se mantiene incluso después de aplicar la ponderación poblacional por `fnlwgt`, confirmando que refleja una desigualdad estructural real. El análisis reveló además que las mujeres presentan mayor concentración ocupacional (índice 0.1590 vs 0.1117 para hombres), sugiriendo patrones de segregación horizontal que contribuyen a las brechas salariales observadas.

Desigualdades por Raza: Los resultados muestran patrones complejos de discriminación racial, donde grupos minoritarios como Asian-Pac-Islander alcanzan proporciones relativamente altas de ingresos superiores (26.56%), aunque con marcadas diferencias por género. La intersección entre género y raza

revela diferencias extremas: hombres Asian-Pac-Islander alcanzan 34.5% de ingresos altos, mientras que mujeres de otras razas reportan proporciones tan bajas como 5.1%, evidenciando un efecto compuesto de discriminación.

Desigualdades por Educación: La educación emerge como un factor determinante en la movilidad económica, con diferencias de hasta 74 puntos porcentuales entre niveles educativos. Individuos con Doctorate alcanzan 74.09% de ingresos altos, contrastando con niveles educativos básicos que raramente superan el 4%. Esta jerarquía educativa refleja la estructura meritocrática del mercado laboral estadounidense, donde el capital humano formal determina significativamente las oportunidades económicas.

Comparación Metodológica: K-means vs Distancia de Gower

La evaluación comparativa de los enfoques de clustering implementados revela trade-offs fundamentales entre eficiencia computacional y capacidad de capturar complejidad socioeconómica. K-means ofrece ventajas claras en velocidad de ejecución (7-40x más rápido) y produce clusters más balanceados con mejor Silhouette Score. Sin embargo, su limitación fundamental radica en el análisis exclusivo de variables numéricas, excluyendo aproximadamente el 60% de la información demográfica disponible.

Por contraste, el clustering jerárquico con distancia de Gower demuestra superioridad en el contexto específico de análisis de desigualdades sociales. Con una separación de ingresos del 99.4% frente al 41.2% de K-means, el enfoque con Gower logra identificar estratos socioeconómicos que reflejan diferencias reales en las oportunidades económicas. Los clusters generados mediante esta metodología muestran patrones interpretables de segregación ocupacional, brechas de género y estratificación racial que permanecen ocultos cuando se utilizan únicamente variables numéricas.

Implicaciones Metodológicas

Los resultados del presente trabajo tienen implicaciones metodológicas importantes para el análisis de desigualdades sociales mediante técnicas de clustering. En primer lugar, demuestran que los datasets socioeconómicos con variables mixtas requieren enfoques especializados como la distancia de Gower, que pueden manejar simultáneamente información categórica y numérica sin pérdida de información relevante. La exclusión de variables categóricas, como ocurre con K-means aplicado sobre variables numéricas puras, resulta en agrupaciones que carecen de relevancia social y que no capturan las estructuras de desigualdad subyacentes.

En segundo lugar, el análisis revela la importancia de considerar la interpretabilidad sociológica sobre la optimización matemática pura. La decisión de utilizar linkage='complete' en lugar de 'average', aunque produjo métricas numéricas inferiores, permitió identificar clusters con significado social claro y aplicación práctica en el análisis de desigualdades. Este trade-off es fundamental cuando el objetivo trasciende la optimización de métricas abstractas para abordar fenómenos sociales complejos.

Finalmente, el trabajo confirma la necesidad de incorporar análisis ponderados mediante variables de peso muestral como `fnlwgt` en estudios poblacionales. Las diferencias identificadas entre análisis ponderados y no ponderados, aunque sutiles en magnitud, tienen implicaciones importantes para la validez externa de los resultados y su generalización a la población de referencia.

Limitaciones y Aportes del Trabajo

El análisis se basa en datos de 1994, por lo que los patrones identificados pueden no reflejar la estructura actual del mercado laboral estadounidense. Además, el carácter transversal de los datos impide establecer relaciones causales firmes sobre los determinantes de las desigualdades observadas.

En síntesis, el presente trabajo demuestra que las técnicas de clustering, cuando se aplican con metodologías apropiadas para datos socioeconómicos complejos, constituyen herramientas valiosas para identificar y caracterizar desigualdades estructurales.