ANÁLISIS DE LA IGUALDAD DE GÉNERO EN LA UE

Máster en Data Science - Nuclio Digital School
Laura Moya, Valentina Santiso, Laia Flores
DSC0222

Índice

1. Introducción

Este trabajo se desarrolla con el objetivo de realizar un análisis sobre la igualdad de género en los distintos países de la Unión Europea, con un principal foco en determinar qué países alcanzarán el objetivo de igualdad de la Agenda 2030 de la ONU.

Esta Agenda 2030 se creó en 2015, cuando los países que forman parte la ONU acordaron 17 objetivos basados en erradicar la pobreza y las desigualdades, para cumplirse todos en 2030. El quinto objetivo, en el cual hemos basado nuestro trabajo, es el de lograr la igualdad entre los géneros.

Para poder resolver nuestra pregunta, buscamos una base de datos con las variables que influencian la igualdad de género en los países de la Unión Europea y así poder estudiar cómo se relacionan entre ellos y ser capaces de predecir qué pasará en 2030.

2. Extracción de datos y análisis del dataset

Los datos los obtenemos del *European Institute for Gender Equality (EIGE),* que nos ofrece los datasets de todos los países de la Unión Europea, separados por hombres y mujeres, para cada año desde el 2013 al 2021, con la excepción de 2014, 2016 y 2018. Considerando estos como valores nulos, tratamos de obtener el valor que más coherencia tiene en el dataset y no modifica la tendencia de este. Asimismo, los añadimos haciendo una media de los años anteriores y posteriores.

Solventados los nulos, juntamos todos los datasets en uno solo, obteniendo finalmente 35 columnas y 504 filas, com una mayoría de variables de tipo *float* y dos categóricas (*Country* y *Sex)*. Las variables son las siguientes:

Variable	**¿Qué representa?**
fulltime_employment_rate	FTE employment rate (%, 15+ population)
duration_working_life	Duration of working life (years, 15+ population)
employed_in_education_health_socialactivities	Employed people in Education, Human Health and Social Work activities (%, 15+ employed)
flexibility_at_work	Ability to take an hour or two off during working hours to take care of personal or family matters (%, 15+ workers)
career_prospects	Career Prospects Index (points, 0-100)
mean_monthly_earnings	Mean monthly earnings (PPS, working population)
mean_equivalised_income	Mean equivalised net income (PPS, 16+ population)
risk_of_poverty	At-risk-of-poverty rate (%, 16+ population)
S20/S80_income	S20/S80 income quintile share (16+ population)

tertiary_education_graduates	Graduates of tertiary education (%, 15+ population)
people_in_education	People participating in formal or non-formal education and training (%, 15+ population)
tertiary_students_education_health_art_field	Tertiary students in the fields of Education, Health and Welfare, Humanities and Art (tertiary students) (%, 15+ population)
caring_kids_eldery_people_everyday	People caring for and educating their children or grandchildren, elderly or people with disabilities, every day (% 18+ population)
cooking_housework_everyday	People doing cooking and/or housework, every day (%, 18+ population)
doing_sport_cultural_activities	Workers doing sporting, cultural or leisure activities outside of their home, at least daily or several times a week (%, 15+ workers)
voluntary_activities	Workers involved in voluntary or charitable activities, at least once a month (%, 15+ workers)
share_ministers	Share of ministers (%)
share_members_parliament	Share of members of parliament (%)
share_members_regional_assemblies	Share of members of regional assemblies (%)
share_board_members_largest_companies	Share of members of boards in largest quoted companies, supervisory board or board of directors (%)
share_board_members_central_bank	Share of board members of central bank (%)
share_board_member_research_org	Share of board members of research funding organisations (%)
share_board_member_public_org	Share of board members of publically owned broadcasting organisations (%)
share_decision_making_body_olympic_org	Share of members of highest decision making body of the national Olympic sport organisations (%)
good_self-perceived_health	Self-perceived health, good or very good (%, 16+ population)
life_expectancy	Life expectancy in absolute value at birth (years)
healthy_life_years	Healthy life years in absolute value at birth (years)
no_smoking_nor_harmful_drinking	People who don't smoke and are not involved in harmful drinking (%, 16+ population)
doing_sport_eating_healthy	People doing physical activities and/or consuming fruits and vegetables (%, 16+ population)
no_medical_examination	Population with unmet needs for medical examination (%, 16+ population)
no_dental_examination	People with unmet needs for dental examination (%, 16+ population)

Sin más nulos ni duplicados en el dataset, evaluamos los *outliers*. La mayoría de variables no presentan ningún comportamiento irregular, a excepción de dos variables: *risk_of_poverty* y *S20/S80_income* (la relación del 20% de la población más rica con el 20% de la población más pobre), sólo en el año 2015. Decidimos sustituirlos por los resultados de 2013, el año real más cercano. Otro preprocesamiento que realizamos es eliminar el valor que representa la media de la Unión Europea, para obtener solo los datos de los países. Con esto se reduce el dataset a 486 filas.

Una vez analizado nuestro dataset buscamos una variable que nos cuantifique cómo está cada país en cuanto igualdad y encontramos el índice de igualdad de género por países, creado por EIGE. El índice puntúa a los miembros de la UE en cuanto a la igualdad de género, utilizando una escala del 1 al 100, donde 1 representa una plena desigualdad y el 100 una total igualdad. Lo añadimos como *Score* en nuestro dataset.

El índice está basado en las diferencias entre hombre y mujer en seis aspectos: *time*, *power*, *health*, *work*, *money* y *knowledge*. Además, también se estudia *violence* y *intersecting inequalities* (desigualdad de género combinado con edad, discapacidad, país de nacimiento, etc.), que deberían influenciar al *score* pero no están contabilizados en este, ya que no hay datos suficientes.

En relación a nuestro dataset, vimos que podíamos separar las variables en estos aspectos también. Por ejemplo, tasa de desempleo con *work*; variables relacionadas con la pobreza y los sueldos con *money*; las de los graduados y estudiantes de diferentes sectores, con *knowledge*; las variables sobre las tareas de casa, el tiempo que se destina al cuidado de los niños y la gente mayor, a *time*; las variables relacionadas con posiciones de poder en política, empresas y bancos, a *power* y las variables relacionadas con salud como esperanza de vida y años de vida saludables, a *health.*

Analizando todas estas variables comentadas, vimos una clara desigualdad entre hombres y mujeres, pero nos faltaba poder justificarlo con nuestros objetivos. *(Ver el notebook de "Igualdad" para observar el análisis univariable realizado).*

3. Objetivo 1: ¿Qué variables son las que más discriminan entre sexos?

Con el objetivo de determinar cuáles son las variables que más discriminan entre hombres y mujeres, desarrollamos un modelo de clasificación. *(Ver el notebook de "Objetivo 1: Modelo de Clasificación").*

Establecemos como *target* la variable sexo, siendo un modelo de clasificación binaria (hombre: 0 y mujer: 1). Después de un análisis de este, viendo que es un dataset balanceado, procedemos a analizar las correlaciones entre variables, donde encontramos un dataset muy correlacionado entre sí, con 41 correlaciones de más del 75%. Además, esto lo podemos ver representado con el cálculo del R2 (calculado más adelante en el objetivo 3 donde se desarrolla un modelo de regresión) que es del 99,14%.

Una vez realizado este análisis, procedemos al modelado. Primero dividimos el dataset en *train* y *test*, en un 70% - 30% respectivamente, y le aplicamos un modelo de XGBoost.

Los resultados obtenidos son estremecedores: un *accuracy* del 100%. Con esto podemos concluir

que existe tal diferencia entre hombres y mujeres que el modelo predice sin error alguno si es un sexo u otro.

A partir de este modelo, analizamos la importancia de las variables al hacer dicha predicción. Asimismo, estudiando tanto el *feature importance* como el *shap values*, determinamos que las variables que más determinan si eres hombre o mujer son:

- Employed_education_health_socialactivities
- Caring_kids_eldery_people_everyday
- Mean_monthly_earnings
- Share_ministers

Con el objetivo de seguir analizando las variables que más influencian, decidimos reducir el dataset a los 6 aspectos generales comentados anteriormente: *time*, *power*, *health*, *work*, *money* y *knowledge*. Para ello, aplicamos la técnica de reducción de la dimensionalidad, en este caso, un *Factor Analysis*.

Decidimos aplicar este método y no el *PCA* estudiado en clase, ya que con este conseguimos reducir el dataset maximizando la varianza, pero no podemos determinar de una manera más o menos clara, qué aspectos representan cada componente obtenido. Igualmente, este ha sido calculado y concluimos que el nombre idóneo de componentes son 2 y que con estos mantenemos el 99,98% de la varianza original.

Siguiendo con el propósito comentado, lo que nos interesa es saber qué contienen estas nuevas variables creadas. Es por esto que aplicamos el *Factor Analysis*. Conseguimos reducir el dataset de 35 variables a 6, manteniendo un 75,45% de la varianza original. Además, calculando los *loadings*, podemos determinar qué aspecto representa cada factor y, con ello, realizar un análisis de cada uno y compararlo entre mujeres y hombres. Concluimos que es en el aspecto de *Power* donde existe una mayor desigualdad entre sexos, teniendo los hombres una mayor representación en órganos de poder, tanto políticos como empresariales. *(Ver el notebook de "Objetivo 1: Reducción de la dimensionalidad").*

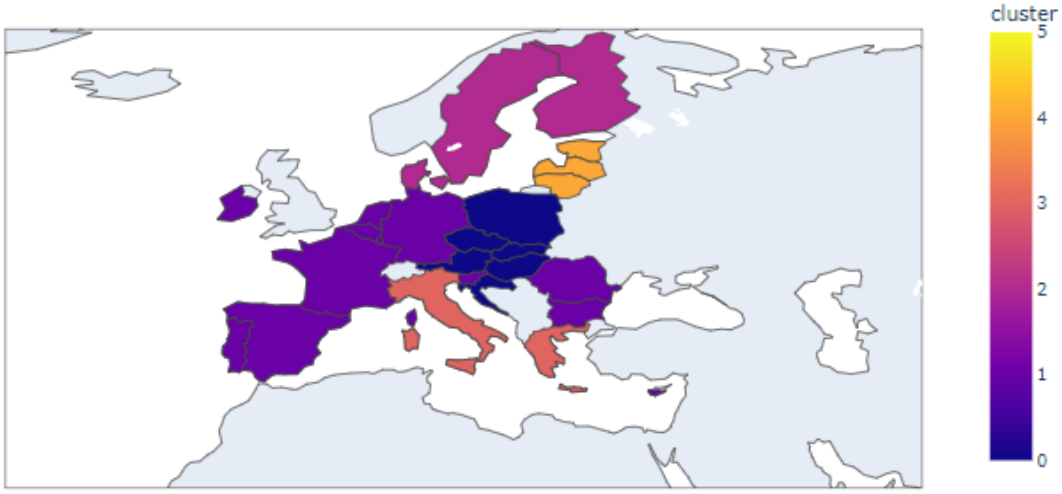
4. Objetivo 2: ¿Qué países presentan unas características similares en cuanto a la desigualdad?

Para responder a esta pregunta desarrollamos un modelo de *clustering*, aplicando el modelo de *K-means* para definir las diferentes agrupaciones de países y que sea más fácil el análisis. *(Ver el notebook de "Objetivo 2: Clustering").*

Para ello, tomamos el dataset resultante de la aplicación de la reducción de la dimensionalidad con el *Factor Analysis*. Primero generamos una nueva estructura con los países en las filas y las *features* en las columnas y, además, calculamos la diferencia entre hombres y mujeres.

Con este dataset modificado, creamos nuevas *features* que utilizaremos en el *K-means*: la media de los últimos 3 años, el valor mínimo y máximo de toda la serie temporal y la pendiente. Con todo esto, aplicando el modelo, identificamos mediante la gráfica del codo el número de clusters a utilizar, obteniendo una K=5.

Con este resultado de K, determinamos los 5 clústers y lo plasmamos en un mapa.



En este podemos visualizar que tiene coherencia geográfica, ya que agrupa países vecinos como España, Portugal y Francia o Finlandia, Dinamarca y Suecia.

Además, para ver si realmente tiene un resultado consistente, analizamos cómo varían estos clusters según cada topic (Time vs Work, Knowledge vs Power, y asi cada topic entre sí). Con este análisis concluimos que cada cluster está bien identificado y no se superponen entre ellos, es decir, cada grupo se comporta de una manera diferenciada en relación a las variables estudiadas. Con esto concluimos que los países están bien clasificados en los clústers..

Una vez identificados estos grupos, estudiamos cómo estos se comportan de acuerdo a la desigualdad, analizando los 6 topics comentados anteriormente.

En general se observa que en todos los clusters existe una clara desigualdad en *Power*. Aun así, cada cluster destaca en algún otro ámbito:

- Cluster 0: existe una mayor desigualdad en *Work*.
- Cluster 1: hay una desigualdad casi nula en *Health* y *Time*.
- Cluster 2: destaca la desigualdad en *Health,* relacionado con la mutilación genital femenina, y una desigualdad casi nula en *Work*.

- Cluster 3: mayor desigualdad en *Time*, debido a un mayor tiempo dedicado al cuidado de niños y adultos y a las tareas del hogar.
- Cluster 4: cluster con mayor desigualdad, debido a una mayor diferencia tanto en cuidado y tareas del hogar (*Time)* y en proyecciones de carrera y trabajos fulltime (*Work)*.

5. Objetivo 3: ¿Cómo se calcula el índice de igualdad de género? Modelo de Regresión

En el tercer objetivo nos planteamos la posibilidad de determinar una fórmula que calcule el *score* de igualdad de género, aplicando un modelo de regresión y calculando los coeficientes que se derivan de este modelo. *(Ver el notebook de "Objetivo 3: Modelo de Regresión").*

Asimismo, como primera instancia, tomamos el dataset inicial (no el reducido) y definimos como target el *Score*, que es el valor que queremos predecir. A este, después de separarlo entre *train* y *test* en un 70% - 30%, respectivamente, aplicamos los diferentes modelos de regresión.

Entre estos, el de *Ridge Regression* es el que da un menor RMSE y, por lo tanto, lo cogemos como válido. Además, para analizar el error, realizamos una gráfica comparando los valores reales contra los obtenidos en la predicción y vemos que la recta de regresión ideal se ajusta consistentemente a nuestra regresión. A su vez, también estudiamos la curva del error, observando que sigue una distribución normal por lo que valida lo concluido. Vemos, por lo tanto, que es un buen modelo y se puede tomar como válido.

Una vez el modelo está definido, procedemos a hacer el cálculo de los coeficientes para la fórmula del *score.* En este observamos que el país tiene una importancia relevante en el cálculo del índice de igualdad de género, ya que los países con una mayor igualdad, tienen un peso mayor y viceversa. Esto tiene sentido, ya que no son sólo las variables que se estudian en este dataset las que determinan esta igualdad, si no que otros factores relacionados con el estado de bienestar de cada país y de desarrollo socioeconómico, que afectan de una manera indirecta a este cálculo.

6. Objetivo 4: ¿Cuál será la situación en 2030? ¿Algún país alcanzará la igualdad de género? Series temporales

Para realizar una predicción de los *scores* aplicamos un modelo de series temporales. Lo primero que hacemos es un *groupby* del *score* y los países, ya que lo que nos interesa aquí es predecir este índice de igualdad de género. *(Ver el notebook de "Objetivo 4: Modelo Prophet de series temporales").*

Para calcular los años del 2022 al 2030 utilizamos el *prophet model*, un modelo que nos permite hacer buenas predicciones, con poco procesamiento de las variables (solo tratamos el score y no generamos nuevas variables) y, además, es un modelo capaz de predecir distintos periodos a largo

plazo de una sola vez.

Probamos de hacer un modelo de Arima, uno de Autoarima, pero teníamos que hacer un preprocesado con lags y no teníamos datos suficientes, además de solo tener una variable y tener que ir país por país generando nuevas. No nos dio un buen resultado, aunque si hubiéramos tenido que predecir solo un año nos hubiera servido, pero no era nuestro caso. Probamos también de hacer un modelo de regresión pero tampoco nos convencieron los resultados y finalmente nos decantamos por el *prophet model*.

Los resultados obtenidos son una predicción del *score* desde el año 2021 al 2030. Teniendo los datos reales de este primer año, podemos calcular el error entre el *score* real de 2021 y la predicción hecha por el modelo

Con las predicciones por cada país, calculamos el RMSE de cada uno y la media entre todos, para analizar el RMSE del modelo y concluir si realiza buenas predicciones o no.

En cuanto a estas, tuvimos distintos resultados por país. Por ejemplo, en Alemania nos salió una predicción que el score aumenta y no tiene cambios sustanciales, y un RMSE bajo de 0,05 comparado con otros países.

En cambio Holanda tiene un comportamiento inusual ya que replica los años anteriores, al tener pocos datos una bajada que tuvo en 2019 hace que el modelo lo replique los siguientes años y nos dio un RMSE más alto.

Otro ejemplo sería España que nos salió que tenía un aumento de hasta 92 de score con un RMSE de 2.26, también por tener una subida sustancial del score los años anteriores.

A pesar de los pocos datos del modelo nos da un RMSE medio de 1.22 que es pequeño al compararlo con la media del score de 65,21, dándonos un 1,86% de la división de la media del score real y el RMSE medio.

Vemos que aunque se trata de un RMSE pequeño, nuestro modelo no es lo suficientemente fiable ya que tratamos con pocos datos.

Con nuestros datos reales, podemos apreciar el impacto que tiene el covid sobre el score en nuestro países, ya que en casi todos se observa una bajada del score en 2019. Buscamos datos del impacto del covid sobre la igualdad de género y encontramos que el covid provocó en las mujeres más estrés laboral y más carga familiar a causa del teletrabajo.

También que según un estudio del Eige Research, las mujeres entre 15-24 años son las que más

perdieron su trabajo debido al covid-19 y en el verano de 2020 los hombres recuperaron 1.4 millones de puestos de trabajo y las mujeres solo 0.7 millones.

Y por último, las mujeres estaban más expuestas al covid debido a sus trabajos en sectores esenciales (salud, cuidados de gente mayor, servicios de limpieza etc,) estaban más en primera línea.

7. Conclusiones:

Finalmente, analizando objetivo por objetivo concluimos:

Gracias a un modelo de clasificación y al estudio del *feature importance*, conseguimos ver qué variables son las que más discriminan entre sexos: Employed_education_health_socialactivities, Caring_kids_eldery_people_everyday, Mean_monthly_earnings y Share_ministers. Por una parte las mujeres son las que más trabajos tienen relacionados con educación, sanidad y arte, y también son las que más cuidan de los niños y la gente mayor. En cambio los hombres son los que más cobran de aquí la brecha salarial y son los que tienen más posiciones de poder en el ministerio.

Para nuestro segundo objetivo hemos visto que había cinco clusters diferenciados, es decir que podemos dividir a los países de la UE en cinco grupos. Cada uno con características diferentes en cuanto a la igualdad, que podemos relacionar con noticias de la actualidad y así demostrar por qué el modelo clasifica a los países en estos clusters.

Con el tercer objetivo, el modelo de regresión, nos permite averiguar qué peso tiene cada variable para el cálculo del score, y poder ver cómo los países tienen también una influencia para el cálculo.

Y finalmente, con el cuarto objetivo vemos que según nuestra predicción, ningún país alcanzará la igualdad de género y por lo tanto es necesario aplicar políticas que reduzcan la brecha entre sexos.

En el 2022, seguimos teniendo claras desigualdades entre hombres y mujeres. La tasa de empleo es mayor en los hombres, aunque las mujeres tienen más formación profesional. Las mujeres dedican más tiempo a los cuidados del hogar y la familia y los hombres siguen cobrando más que las mujeres en posiciones de trabajo similares. Además son los hombres los que tienen más representación en puestos de trabajo con posiciones de poder. En definitiva, es necesario que se sigan aplicando políticas para poder llegar a una igualdad de género, los datos ya hablan por sí solos.