Facultad de Ciencias Empresariales

APLICACIÓN DE LAS TÉCNICAS DE BUSINESS ANALYTICS PARA EL ESTUDIO DE LA EDUCACIÓN DE CALIDAD Y SU POSIBLE RELACIÓN CON LA EFICACIA MERITOCRÁTICA

Autor: Pelayo Plaza Jorge

Director: José Ramón Vallejo Rodrigo

MADRID | Abril de 2026

Índice

DECLARACIÓN DE USO RESPONSABLE DE LA IA v

RESUMEN vi

INTRODUCCIÓN 1

Interés del Estudio 1

Fines y Objetivos del TFG 3

Metodología 3

Estado del arte 4

DEFINICIÓN DE LAS FASES DEL PROYECTO 7

Arquitectura de los datos 7

Origen de los datos 10

Almacenamiento de los datos 11

Tratamiento de los datos de entrada 11

Análisis exploratorio de los datos de entrada 14

Tratamiento de los datos de salida 22

Análisis exploratorio de los datos de salida 23

DETALLE TÉCNICO DEL PROYECTO 29

Tratamiento de los datos para la generación del modelo 29

Clustering jerárquico 30

Sistema de scores 31

Inserción de las variables de salida 37

Resultados finales 39

CONCLUSIONES 45

Interpretación de resultados 45

Limitaciones y próximos pasos 46

BIBLIOGRAFÍA 47

ANEXOS 49

ÍNDICE DE GRÁFICOS Y TABLAS

Tabla 1: Código de los indicadores de infraestructura escolar 13

Tabla 2: Histograma de distribución de la variable escs_trend 16

Tabla 3:Boxplot de la variable Proportion of schools with access to the internet for pedagogical purposes [Lower secondary] 17

Tabla 4: Boxplot por país de la variable Becas 18

Tabla 5: Matriz de correlación de las variables de entrada..... 20

Tabla 6: Mapa de correlación de las variables de entrada corregido para evitar la multicolinealidad 21

Tabla 7: Histograma de inserción Laboral..... 25

Tabla 8: Boxplot de la brecha salarial 26

Tabla 9: Boxplot por país de la movilidad rural 27

Tabla 10: Mapa de correlación de las variables de salida..... 27

Tabla 11: Clustering jerárquico de países en función de las variables de educación de calidad (entrada) 30

Tabla 12: Países con mayores mejoras relativas en calidad educativa 32

Tabla 13: Código para la implementación del Random Forest 33

Tabla 14: Importancia de las variables en la definición de clústeres..... 34

Tabla 15: Comparación de la puntuación de calidad educativa de España, Francia, Reino Unido, Italia, Alemania, Grecia y Portugal 35

Tabla 16: Mapa de calor de los scores educativos..... 36

Tabla 17: Carga del dataset de salida y adecuación de nombres de paises al sistema ISO de la OCDE..... 37

Tabla 18: 10 países con mejores scores de salida..... 38

Tabla 19: Matriz de eficiencia meritocrática 40

Tabla 20: Ranking de países según el score global 42

Tabla 21: Peso de las variables de entrada en las variables de salida..... 43

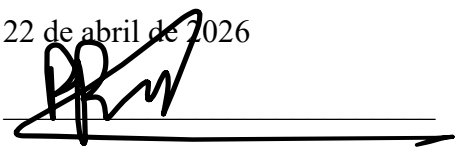
DECLARACIÓN DE USO RESPONSABLE DE LA IA

Por la presente, yo, Pelayo Plaza Jorge, estudiante de E3 Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Aplicación De Las Técnicas De Business Analytics Para El Estudio De La Educación De Calidad Y Su Posible Relación Con La Eficacia Meritocrática", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
2. **Interpretador de código:** Para realizar análisis de datos preliminares.
3. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
4. **Generador previo de diagramas de flujo y contenido:** Para esbozar diagramas iniciales.
5. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 22 de abril de 2026

Firma: 

RESUMEN

Este trabajo analiza la relación entre la educación de calidad y la eficacia meritocrática mediante la aplicación de técnicas de *Business Analytics* sobre datos internacionales procedentes de organismos como Eurostat, OCDE y PISA. A partir de la construcción de dos conjuntos de variables: uno de entrada, vinculado a factores educativos (infraestructura, gasto, becas, calidad docente y contexto socioeconómico), y otro de salida, relacionado con resultados en el mercado laboral (inserción, brecha salarial, sobre cualificación y movilidad), se desarrolla un modelo cuantitativo que pretende evaluar la eficiencia de los sistemas educativos en distintos países europeos. Mediante técnicas como el *clustering jerárquico*, sistemas de puntuación y modelos de *Random Forest*, se intenta identificar patrones estructurales y se observan resultados sobre los países en función de su capacidad para transformar inversión educativa en resultados meritocráticos. Los resultados parecen mostrar una fuerte influencia del gasto educativo y del contexto socioeconómico en los resultados laborales, así como el papel estratégico de las becas como mecanismo de eficiencia. Asimismo, se intuyen diferencias geográficas, destacando casos de alta eficiencia y otros de posible desconexión estructural entre educación y mercado laboral. El estudio parece indicar que la relación entre educación y la eficacia meritocrática es compleja y está mediada por factores estructurales, lo que cuestiona la existencia de una igualdad de oportunidades plena en los sistemas analizados.

INTRODUCCIÓN

Interés del Estudio

Ante el panorama de un mundo sobre acelerado, con cambios vertiginosos y una evolución social y tecnológica que se hace para muchos casi ininteligible, acecha sobre nosotros una abrumadora sensación de que gran parte de la sociedad va a quedarse atrás. Esto explica el auge en los últimos años de la implementación de medidas dirigidas a mejorar la *justicia social*, como los Objetivos de Desarrollo Sostenible, y la gran cantidad de obras que buscan concienciar a la sociedad sobre la gravedad de un avance sociocultural que no incluya a todos sus ciudadanos. En este sentido, textos como *Laudato Si',* la encíclica del Papa Francisco donde se aboga por eliminar la *cultura del descarte*, nos ayudan a tener una visión global del problema al que nos enfrentamos actualmente. La única espada que parece poder esgrimirse en esta tesitura es que el avance técnico vaya aparejado de un avance en la educación y la cultura orientado al debate y el juicio crítico sobre la bondad de estos avances. Volviendo a Laudato Sí': "Entre los componentes sociales del cambio global se incluyen los efectos laborales de algunas innovaciones tecnológicas, la exclusión social, la inequidad en la disponibilidad y el consumo de energía y de otros servicios, la fragmentación social, el crecimiento de la violencia y el surgimiento de nuevas formas de agresividad social, el narcotráfico y el consumo creciente de drogas entre los más jóvenes, la pérdida de identidad." (Laudato Sí', §46) "Esto nos exige un esfuerzo para que esos medios se traduzcan en un nuevo desarrollo cultural de la humanidad y no en un deterioro de su riqueza más profunda" (Laudato Sí', §47)

Es esta preocupación por el desarrollo educativo y cultural que nos lleva a abordar este problema desde el punto de vista de la educación de calidad y la igualdad de oportunidades, haciendo especial hincapié en el concepto de la meritocracia. No obstante, estos tres conceptos se han asentado en el discurso público sin muchas veces tener una definición concreta de qué significan, por lo que debemos, antes de continuar, establecer a qué nos referimos cuando hablamos de ellos.

La educación de calidad como concepto ha generado no pocos debates y ha sido causa de un gran número de investigaciones y artículos académicos. Si bien no existe un consenso

que certifique una definición clara, creemos que la que brinda Carla Figueiredo, profesora de la Universidad de Porto, establece un marco suficientemente claro y preciso por el que nos podemos guiar: «La educación de calidad está orientada a preparar a las personas para que contribuyan al proyecto social y político, adquiriendo conocimientos, habilidades y competencias para ser productivos; para ello se necesitan profesores y líderes motivados y bien preparados, condiciones adecuadas y un estrecho seguimiento y evaluación» (Figueiredo, 2025).

La igualdad de oportunidades la definen Roemer (profesor en la Universidad de Yale) y Trannoy (profesor en la Universidad de Aix-Marseille) como la «búsqueda de compensar la diferencia en los resultados atribuibles a la suerte, pero no aquellas diferencias en los resultados de las que los individuos son responsables» (Roemer et al., 2016).

Si bien el término meritocracia nace de Michael Young en 1958, la definición que aporta Chang-Hee Kim delimita de forma sublime y muy ilustrativa el concepto: «La meritocracia es un sistema social en el que el avance en la sociedad se basa en las capacidades y méritos individuales, en vez de en el origen familiar, la riqueza o el origen social» (Chang-Hee Kim et al., 2017).

Una vez delimitado el marco teórico, observamos que en la práctica se ha establecido en la dialéctica social que cuanto más se esfuerce uno, más lejos podrá llegar, y que no existen puertas cerradas para nadie. Se impone así una lógica de desplazamiento social completamente elástico, que nos equipara a todos y nos permite empezar desde un mismo punto de partida. Este sistema es el que se conoce hoy como meritocrático. Si bien no podemos negar ejemplos claros donde los sucesos han seguido este cauce (Amancio Ortega, levantando un imperio textil desde cero en Arteixo), la mutación social que acaece en estos tiempos, el síndrome de nuestra evolución desenfrenada, ha difuminado las líneas de este concepto. En los últimos años, el frente político y económico ha seguido defendiendo la meritocracia como un sistema confiable y vigente, mientras que el artístico ha sido cada vez más crítico con ella. En este ámbito merece una mención especial el libro *La Tiranía del Mérito ¿Qué ha sido del bien común?* del Profesor de Harvard Michael J. Sandel, cuyo libro ha servido de gran inspiración para el desarrollo de este trabajo. Por su parte, autores como David Foster Wallace o cineastas como Sean Baker

han centrado gran parte de su obra en desmontar lo que ellos consideran el *mito* del *sueño americano,* exponiendo una situación de grave desigualdad en los Estados Unidos.

En lo que concierne al viejo continente, la Unión Europea y sus Estados Miembros han adoptado medidas de justicia social más agresivas que sus homólogos americanos. Parecen haberse alejado del *laissez faire* americano, buscando paliar los efectos nocivos de la evolución social. De este modo, es interesante hacer un estudio comparativo de ambas formas de administración del cambio social, cultural y tecnológico, para intentar buscar los motivos que nos han traído hasta aquí y las posibles soluciones al problema, si es que existen.

Fines y Objetivos del TFG

El objetivo de este trabajo sería, por tanto, intentar aproximarse a la identificación de la situación social actual en relación con la educación y la igualdad de oportunidades, así como explorar posibles medidas que podrían contribuir a reducir la desigualdad. De esta forma, se trataría de analizar hasta qué punto podría considerarse que nos regimos por un sistema en el que el mérito resultaría suficiente para alcanzar puestos de responsabilidad o si, por el contrario, el esfuerzo personal no bastaría para lograr determinados objetivos. En cualquier caso, este análisis se plantearía como un acercamiento, ya que no sería posible determinar de forma concluyente si un país puede considerarse plenamente meritocrático o no. Del mismo modo, se estudiaría el papel de la educación en este contexto y cómo podría influir en la configuración de una línea de partida más equitativa para el conjunto de la ciudadanía.

Metodología

Para llevar a cabo este estudio, identificaremos bases de datos que contengan información relativa a distintos baremos que pueden ser sintomáticos del *status quo*. Factores como los índices de desigualdad, los sesgos de representatividad, la movilidad social e intergeneracional, y los ránquines educativos pueden servir para dilucidar el objetivo del trabajo. Una vez dibujado un boceto de la situación actual, propondremos distintos

escenarios variando los factores, estableciendo hipótesis que puedan ser de utilidad para hacer un análisis de la situación desde distintos puntos de vista.

No obstante, al tratarse la meritocracia de un concepto que puede ser definido desde infinidad de perspectivas, hemos preferido definir un concepto que se adecúe más a la propuesta del trabajo, relacionada con la educación de calidad, de forma que no nos apropiemos en los resultados de una explicación global de la idea, sino únicamente de un proxy de la misma, vinculado al objeto del estudio y con variables que lo definen de forma parcial – frente a la dificultad de hacerlo de forma global y consensuada. Llamaremos a este acercamiento "eficacia meritocrática".

Todo este proceso se llevará a cabo mediante la aplicación de técnicas avanzadas de Data Science y Business Analytics, integrando un flujo de trabajo híbrido que combina la extracción y el preprocesamiento de datos en RStudio (lenguaje R) con el modelado predictivo en Python.

Para el análisis de los datos, se han implementado técnicas de aprendizaje no supervisado, específicamente algoritmos de Clustering, con el fin de *intentar* identificar patrones estructurales y agrupar a los países según sus niveles de inversión y recursos. Posteriormente, se ha utilizado un modelo de aprendizaje supervisado basado en Random Forest para determinar la importancia relativa de cada variable en la clasificación educativa. Finalmente, el proceso culmina con el desarrollo de un modelo de Scoring Inteligente, que *busca* evaluar y predecir el desempeño de los sistemas educativos mediante una ponderación optimizada por algoritmos de inteligencia artificial.

Estado del arte

Haciendo un acercamiento al panorama nacional, los profesores de la Universidad de León, Alejandro López-González y Paz Benito del Pozo, trazan a la perfección una guía de cómo evaluar la desigualdad por grupos de edad y geográficos para espacios determinados. Su estudio sobre la desigualdad en Madrid, Barcelona, Valencia y Sevilla se establece un buen modelo a seguir para cuantificar y analizar la desigualdad en distintas zonas geográficas, sin llegar a ahondar, no obstante, en cómo solucionar ese problema o las causas que nos han llevado hasta él. El uso de métricas cuantitativas y refinamiento

estadístico de los datos, mediante el uso de Z-scores y sesgos de desigualdad traza un buen camino a seguir a la hora de analizar nuestros datos. Por su parte, Leandro Prados de la Escosura, profesor emérito de la Universidad Carlos III de Madrid, y Blanca Sánchez-Alonso, catedrática de Historia Económica de la Universidad San Pablo CEU, hacen un estudio magnífico sobre la relación de la evolución económica histórica en España, relacionada con la evolución de la desigualdad, haciendo unas conclusiones finales que parecen indicar que la implantación del modelo del bienestar desde los años 80 han llevado a una mitigación de la desigualdad en nuestro país. Sin embargo, trabajos llevados a cabo por colectivos orientados a la mejora de la justicia social, como el Colectivo Ioé, parece llegar a la conclusión de que ha habido un claro aumento de la desigualdad en los últimos años, marcado por un decrecimiento del patrimonio y un reparto desigual de la riqueza. Ambos trabajos abordan el tema desde un punto de vista meramente económico, sin contar con otros factores como los que buscamos añadir al análisis.

Ampliando el punto de mira y centrándonos en la situación de Europa, Nosratabadi, Atobishi y Hegedűs estudian las consecuencias de la no inclusión tecnológica de todos los ciudadanos y cómo ésta se traduce en un peor índice de cumplimiento de los Objetivos de Desarrollo Sostenible, pero sin hacer ningún planteamiento sobre la meritocracia. Bandiera, Kotia, Lindenlaub, Moser y Prat, profesores de las universidades de Yale, Columbia, Vrije y LSE, sí que hacen un enfoque sobre la tecnología y la meritocracia, abordándolo, no obstante, desde el punto de vista de la desigualdad de salarios y dejando de lado el apartado educativo.

Toby Napoletano, Doctor en filosofía y profesor de la Universidad de California, Merced, aborda la cuestión desde la misma perspectiva que nosotros pretendemos en este trabajo, usando la dialéctica social como métrica de evaluación de la situación actual. No obstante, creemos que su análisis filosófico sobre la educación meritocrática y la igualdad de oportunidades podría beneficiarse de un análisis cuantitativo, aplicando las técnicas de Business Analytics, además de un análisis hipotético que nos permita analizar casos paralelos con los mismos datos, en vistas de buscar modelos sociales que mejoren la justicia social o que nos sirvan para corroborar que el modelo por el que nos regimos a día de hoy es el más eficiente en la búsqueda de una sociedad igualitaria.

De este modo, buscamos exponer que el trabajo que aquí se realiza es innovador y busca esclarecer la cuestión de estudio acercándose al tema desde una perspectiva distinta a la que anteriormente se ha acudido para abordarlo y que creemos puede aportar información muy interesante para obtener una visión global del problema. Una visión complementaria con las investigaciones mencionadas previamente, y que sirven para analizar el problema de la forma más pormenorizada posible.

DEFINICIÓN DE LAS FASES DEL PROYECTO

Arquitectura de los datos

La primera fase del proyecto consistía en la obtención de datos que fuesen capaces de ligar el carácter etéreo e indeterminado de conceptos como la *educación de calidad* y la *eficacia meritocrática* a una realidad cuantificable a partir de la cual se pudiese hacer un estudio cuantitativo, apoyados por las herramientas de *Business Analytics*.

Para lo cual, establecimos un marco teórico que acotase las variables que relacionan la educación de calidad y la eficacia meritocrática. En una sociedad en la que se busca que le educación sirva como paliativo de las desigualdades y equilibrio de las fuerzas del azar para que todos los ciudadanos dependan únicamente de sus capacidades como timón de su vida laboral, innumerables variables entran en juego. Ante la imposibilidad de encerrar la complejidad del mundo en un modelo, recurrimos a aquellas de mayor peso y que creemos podrán ser lo más ilustrativas posible.

En esta sociedad idílicamente meritocrática debería haber una correlación positiva casi perfecta entre el nivel de habilidades requeridas para desarrollar una actividad laboral concreta y las que tiene el trabajador. Un nivel demasiado alto de sobre cualificación podría entenderse como que la gente que merecería estar en un puesto más alto no lo está por un fallo del sistema a la hora de distribuir los puestos de trabajo más altos entre los mejores cualificados. También podría significar que existe un desequilibrio entre las licenciaturas y el mercado laboral. Otro punto de vista desde el que se puede defender la sobre cualificación es que el sistema educativo es de una inmensa calidad, por lo que no podríamos criticar ni la efectividad del sistema meritocrático ni la educación de calidad, puesto que podría argumentarse que saber de más nunca va en detrimento, si es que el mercado laboral pide unas aptitudes que el sistema educativo supera. Del mismo modo, si existe una gran tasa de infra cualificación, existiría un fallo en la educación de calidad, puesto que esta no es capaz de enseñar a los estudiantes a despeñarse al nivel requerido en el mundo laboral.

Una de las bases de la eficacia meritocrática es que los puestos más altos deben estar asignados a los mejor preparados para desempeñarlo. El problema surge cuando nos hallamos, por ejemplo, ante el altísimo número de licenciados en España, por lo que

debemos acudir al expediente académico. También esta horquilla es de una magnitud tal que no podríamos diferenciar a una persona de otra para saber que es la indicada para ejercer ese puesto. A este efecto, un factor externo, como la suerte, la experiencia o las conexiones son las que nos deben acercar a la solución. Llegados a este punto, las diferencias son tan ínfimas que la única explicación debería ser la suerte. Siguiendo este hilo, las posiciones de más alto prestigio también deberían corresponderse con un salario más alto, el cual sirve como indicativo numérico de la posición en la que se haya cada trabajador. No es extraño para nadie que, en el modelo socioeconómico en el que nos movemos, existe una correlación entre la voluntad de acceder a un puesto y la retribución económica que este otorga.

Una vez establecido el marco, nos cuestionamos cuáles serían las variables de entrada que podrían apoyar la hipótesis de que una educación de calidad desemboque en eficacia meritocrática. Para realizar esta tarea, era indispensable la obtención de datos que pudiesen ser significativos de una educación que, según las definiciones previamente mencionadas, pudiese tildarse *de calidad* y cuyos resultados facilitasen la creación de un marco donde poder discutir la existencia de la supuesta meritocracia que marca el tema principal de este trabajo. De este modo, aunque la enseñanza de calidad no funcione como parámetro único que ratifique la existencia de la eficacia meritocrática, podemos asumir que ésta solo aparecerá si hay un clima propicio de igualdad que establezca, sino homogénea, una línea de partida suficientemente uniforme como para descompensar las desigualdades, y permitir a los individuos progresar en función de sus capacidades y méritos individuales, aplacando los designios de la suerte.

Para llevar a cabo este proceso, la arquitectura de los datos se establece de forma bipartita de la siguiente manera:

En primer lugar, unas variables de entrada sintomáticas de una educación de calidad, véase:

- Infraestructura escolar
- Gasto público por alumno
- Estabilidad o incremento presupuestario
- Calidad del profesorado
- Resultados educativos

En segundo lugar, unas variables de salida que podrían significar una asociación entre una mejor educación y mayor eficacia meritocrática:

- Rendimiento ajustado por contexto
- Inserción laboral por nivel educativo
- Brecha salarial ajustada
- Movilidad social intergeneracional
- Correspondencia educación-empleo

Para poder unir las variables de entrada con las de salida, habremos de construir dos datasets, uno para cada conjunto.

En el primer dataset figurarán variables que se pueden ser sintomáticas de una educación de calidad, o, por lo menos, que pueden servir como proxys de variables que servirían para definirla. El segundo, por su lado, se compondrá de variables que creemos que pueden ser significativas de eficacia meritocrática. Es importante destacar en este aspecto, que no tratamos de analizar una meritocracia a nivel personal, si no, más bien, una especie de "eficacia meritocrática macro". En vez de analizar la eficacia meritocrática teniendo en cuenta el desempeño de sujetos concretos (lo cual nos sería imposible con los datos recopilados), buscamos analizarla a nivel de países. De este modo, el estudio de la significación de nuestra aproximación a la educación de calidad en nuestra aproximación de eficacia meritocrática dependerá del grado de correlación que exista entre los resultados del dataset de variables de entrada y el resultado de las variables del dataset de salida.

De este modo, la relación que queremos establecer es la siguiente: un país con alto nivel de infraestructura escolar, alto gasto en educación, mucha inversión en becas y desempeño educativo desvinculado de factores socioeconómicos, debería traducirse en un rendimiento más ajustado por contexto, mayor inserción laboral en función del grado educativo, una brecha salarial más ajustada, movilidad social más elástica y una alta correspondencia entre nivel educativo y posición laboral.

Para poder analizar esta relación desde una realidad cuantificable, apoyada teóricamente en el marco previamente expuesto, tuvimos que recopilar una serie de datos que no encierran a la perfección el modelo expuesto, pero que, no obstante, pueden servir como proxys del objeto de estudio. Por tanto, cuando hacemos referencia a la educación de

calidad o la eficacia meritocrática, no podemos pretender que se haga un análisis que encierre la realidad universal de los conceptos, si bien sí podemos asumir que recogen datos suficientemente significativos para poder desarrollar la investigación.

Origen de los datos

Estos datos se obtuvieron, principalmente, de las bases de datos de organizaciones supranacionales como son la Unión Europea, vía su portal público de bases de datos *Eurostat*, y la Organización para la Cooperación y el Desarrollo Económicos (OCDE), mediante su portal público de bases de datos *OCDE Data Explorer*.

De *Eurostat* se obtuvieron datasets con información referente a los países de la zona Euro, junto con otros de gran influencia en la economía europea – si bien no forman parte de la zona Euro o la Unión Europea - de los siguientes temas: sobre cualificación por actividad económica; población en hogares privados por nivel de educación y estatus laboral; población en hogares privados por nivel de educación, país de origen y grado de urbanización; tasas de empleo por nivel de educación; medianas de ingresos por hora para todos los empleados (excluyendo aprendices) por sexo.

De *OCDE Data Explorer* se obtuvieron datasets con información referente a países miembros de la organización de los siguientes temas: Educación obligatoria; gasto gubernamental en educación; nivel de cumplimiento del Objetivo de Desarrollo Sostenible 4, referente a la educación de calidad.

Por último, para la información referente a los resultados educativos, se hizo uso de la información recogida en las bases de datos del *Informe PISA*, cuyo objetivo es medir el rendimiento académico de los países en distintas disciplinas educativas, y que pudimos obtener gracias a su dataset de uso público.

La coexistencia de estas fuentes se justifica mediante el uso de estándares internacionales compartidos, como la Clasificación Internacional Normalizada de la Educación (ISCED) y los códigos geográficos, que garantizan la consistencia técnica en la unión de los registros. Es preciso señalar que este proceso de integración conlleva una reducción deliberada de la muestra original hacia una intersección de alta fidelidad. Al realizar el

cruce de variables, se descartan aquellos registros que no presentan una información simultánea en los tres organismos.

Sin embargo, esta pérdida información a nivel geográfico se traduce en una ganancia de robustez metodológica, al crear una base de datos más compacta y con menos posibilidad de generación de ruido en el apartado metodológico. La intersección resultante constituye una muestra representativa asegurando que el modelo de Machine Learning trabaje sobre un conjunto de datos validados y sincronizados.

Almacenamiento de los datos

Los datos se descargaron de las respectivas páginas web. Si bien algunos de ellos se descargaban en formato .csv, aquellos procedentes de *Eurostat* debían accederse mediante la utilización de un comando de descarga, desde la plataforma Rstudio. Una vez todos los datos hubieron sido descargados y almacenados en el script de R, se pasó a hacer una limpieza de los mismos, dejando solo aquella información cuantitativa que resultase necesaria para el proceso de programación posterior. Para agrupar los datos, se utilizaron como criterios principales el país o zona geográfica y el año, para poder hacer un estudio individualizado de cada región, al igual que para tener la opción de procesar y analizar los datos de forma temporal, aportando más profundidad y comprensión al tema de estudio.

Tratamiento de los datos de entrada

A partir de las bases de datos generales, se hicieron particiones para hacer un estudio pormenorizado de las variables de entrada a estudiar.

En un primer lugar, se extrajeron del dataset de ODS 4 una serie de variables con el objetivo de analizar la infraestructura escolar en distintos países, componiéndose de las siguientes columnas: REF_AREA, SDG_INDICATOR, SDG.indicator, SDG_SERIES, SDG.series, TIME_PERIOD y OBS_VALUE.

Para simplificar el dataset, se hizo una limpieza de datos para trabajar únicamente con las variables: REF_AREA, SDG.series, TIME_PERIOD y OBS_VALUE. Con el mismo

objetivo, se cambiaron los nombres de SDG_SERIES, que marcaban los distintos criterios de calidad de infraestructuras, por nombres sistemáticos que facilitasen la lectura del dataset, quedando del siguiente modo:

INDICADOR	CÓDIGO
Proportion of schools with access to electricity [Primary]	S1
Proportion of schools with access to basic drinking water [Primary]	S2
Proportion of schools with access to basic drinking water [Lower secondary]	S3
Proportion of schools with access to electricity [Lower secondary]	S4
Proportion of schools with access to electricity [Upper secondary]	S5
Proportion of schools with access to basic drinking water [Upper secondary]	S6
Proportion of schools with access to the internet for pedagogical purposes [Primary]	S7
Proportion of schools with access to the internet for pedagogical purposes [Lower secondary]	S8
Proportion of schools with access to the internet for pedagogical purposes [Upper secondary]	S9

Proportion of schools with access to adapted infrastructure and materials for students with disabilities [Lower secondary]	S10
Proportion of schools with access to adapted infrastructure and materials for students with disabilities [Primary]	S11
Proportion of schools with access to adapted infrastructure and materials for students with disabilities [Upper secondary]	S12

Tabla 1: Código de los indicadores de infraestructura escolar

Por último, se procedió a cambiar el nombre de la columna OBS_VALUE por "infraestructura", de modo que se diferenciase a la hora de unir todos los datasets en uno definitivo.

En segundo lugar, a partir del dataset de gasto gubernamental en educación, se extrajo un subset de gasto que solo incluyese aquel que computase todos los niveles educativos, de modo que no se examinasen únicamente etapas aisladas, sino que se considerasen todas en conjunto. Para este subset, se escogieron únicamente las columnas REF_AREA, TIME_PERIOD y OBS_VALUE, que, al igual que en infraestructura, recibió posteriormente el nombre de "gasto".

Respecto al indicador 4.c, se obtuvieron los datos relacionados con la cualificación docente, bajo la denominación de calidad_prof. Es necesario acotar que este indicador se refiere específicamente a la proporción de docentes que cuentan con las titulaciones y la formación pedagógica mínima exigida por los estándares nacionales de cada país. Para garantizar la coherencia con el resto del estudio, se aplicó un filtro previo en la variable SEX para seleccionar únicamente los registros que contemplan a ambos sexos, asegurando así una visión agregada de la formación del profesorado en los sistemas analizados.

En cuarto lugar, se trataron los datos provenientes del dataset de PISA, seleccionando las variables *cycle*, *cnt* y *escs_trend*. En cuanto a ésta última, es importante precisar que, más que una medida de rendimiento académico directo, este indicador actúa como un *proxy* del contexto socioeconómico de origen, permitiendo observar la tendencia en la composición del capital social y cultural en el entorno educativo de cada país. Para la posterior integración de las bases de datos, se cambiaron los nombres de las variables *cnt* y *cycle,* por REF_AREA y TIME_PERIOD, respectivamante.

Por último, del dataset ODS 4, se obtuvieron los datos relacionados con la inversión en becas por parte de los estados, siendo este el indicador 4 B. Aunque este indicador no contabiliza la inversión interna total de cada sistema educativo nacional en becas, se utiliza en este modelo como una métrica del apoyo financiero internacional recibido y del esfuerzo en cooperación educativa. Al igual que con la calidad del profesorado, el subset se compuso únicamente de las variables REF_AREA, TIME_PERIOD y flows_becas, siendo flow_becas lo que antes se denominaba OBS_VALUE.

Una vez explorados y tratados todos los dataset por separado, se unieron todos en un dataset definitivo llamado “df”, utilizando como clave de unión las variables REF_AREA y TIME_PERIOD, lo cual conllevó la transposición de las series de infraestructura.

En cuanto a TIME_PERIOD, debido a que los datos PISA únicamente se publicaron en los años 2012, 2015 y 2018, se hizo una limpieza para trabajar únicamente con los datos de esos años.

Análisis exploratorio de los datos de entrada

En una primera instancia se hizo un análisis de valores vacíos, observándose que REF_AREA y TIME_PERIOD no presentan valores faltantes (0 %), lo cual es positivo, ya que ambas variables identifican el país y el año de la observación y son fundamentales para la estructura del conjunto de datos. Asimismo, la variable escs_trend presenta un porcentaje relativamente bajo de valores perdidos (2,05 %), lo que indica que la información sobre la tendencia del estatus socioeconómico del alumnado está prácticamente completa para la mayoría de las observaciones. Por otro lado, varias variables presentan niveles medio de datos faltantes, situándose aproximadamente en

torno al 50 % o más. En concreto, flows_becas tiene un 51,33 % de valores ausentes, calidad_prof un 52,73 % y gasto un 52,20 %.

Una situación similar se observa en los indicadores relacionados con infraestructura escolar básica. Las variables S2, S3, S6, S7, S8 y S9 presentan porcentajes de valores perdidos que oscilan aproximadamente entre el 56 % y el 60 %.

Finalmente, los indicadores S10, S11 y S12, que miden el acceso a infraestructura y materiales adaptados para estudiantes con discapacidad, presentan los niveles más altos de datos faltantes, con un 95,07 % de valores ausentes. Ante este nivel tan elevado de valores vacíos, se decidió prescindir de esas variables, ya que no aportaban apenas información y podrían generar ruido innecesario en el modelo posterior.

No obstante, a diferencia de los indicadores de infraestructura descartados, se determinó que las variables con niveles de valores vacíos cercanos al 50-60% resultan datos analíticos muy importantes para desarrollar el cauce del estudio. La ausencia de estos datos no responde a una falta de relevancia, sino a la periodicidad irregular de las encuestas internacionales y los ciclos de reporte de los organismos oficiales.

De este modo, obtenemos un dataset de entrada final con unas dimensiones de 2.279.805 filas y 12 columnas. Con un total de 75 países miembros de la OCDE y la UE y en un panorama temporal acotado a los años 2012, 2015 y 2018, al ser éstos los únicos reportados en el informe PISA y, por tanto, los únicos años en los que obtendríamos valores completos para cada observación.

Después de hacer un tratamiento de outliers, decidimos tan solo eliminar un valor atípicamente alto en el dataset gasto, ya que parecía referirse más a un error de computación del dataset que un valor real. Por su parte, los indicadores de infraestructura escolar (S2, S3, S6, S7, S8 y S9) sitúan sus valores en torno a 100, ya que tanto el primer cuartil como la mediana y el tercer cuartil coinciden frecuentemente en este valor. Esto indica que en muchos países el acceso a estas infraestructuras es prácticamente universal. No obstante, los valores mínimos, que en algunos casos descienden hasta aproximadamente 38 o 55, reflejan la existencia de países o contextos específicos donde el acceso es considerablemente menor. Aunque estos valores no necesariamente constituyen outliers extremos, sí muestran diferencias relevantes entre países. Se ha preferido mantener estos valores tan alejados de sus respectivas medias y medianas ya

que estos valores aportan más información, al mostrar realmente dónde se encuentran las diferencias entre los distintos países en cuanto a infraestructura escolar. Si bien la mayoría de zonas geográficas tendrán, por ejemplo, cerca de un 100% de escuelas de educación primaria con acceso a agua potable, aquellos países que no lleguen a ese porcentaje se verán severamente penalizados, aportando información muy importante a la hora de extraer conclusiones.

Después de llevar a cabo estos pasos, se procedió a realizar una serie de gráficas que explicasen la distribución de cada una de las variables. Se realizaron: histogramas, boxplots, boxplots por país, boxplots por año y una matriz de correlación. A continuación, se hará un análisis de las gráficas más destacables:

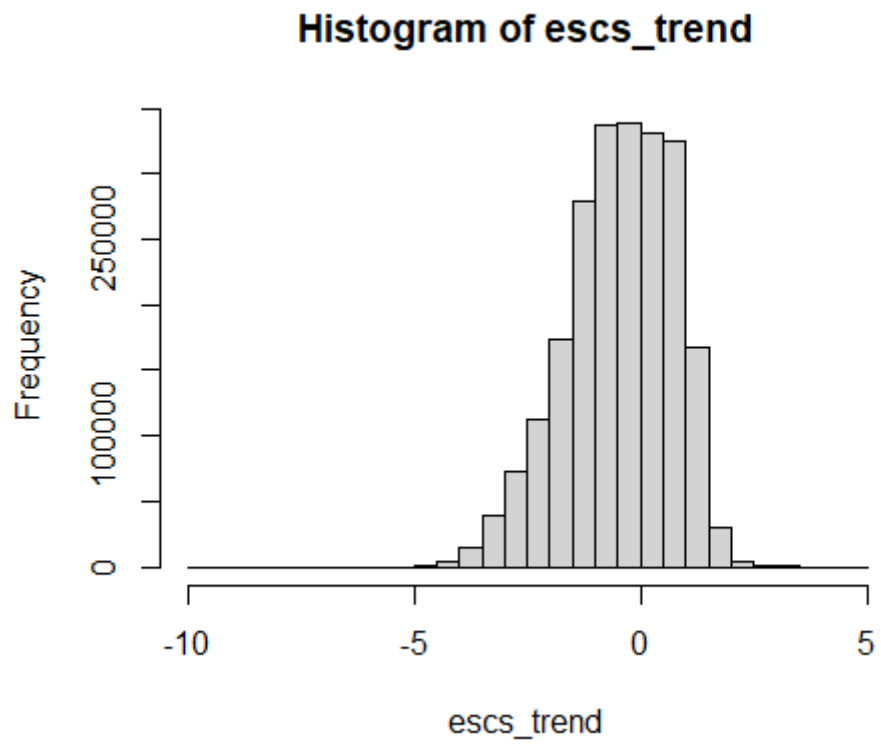


Tabla 2: Histograma de distribución de la variable escs_trend

La gráfica muestra el histograma de la variable *escs_trend*, que representa la tendencia del estatus socioeconómico del alumnado. A partir de la distribución observada, se aprecia que la mayoría de las observaciones se concentra alrededor de valores cercanos a 0, lo que indica que, en general, los cambios en esta tendencia no son muy pronunciados. La

forma del histograma es aproximadamente unimodal y cercana a una distribución normal, con un pico central bien definido y una disminución gradual de la frecuencia hacia ambos extremos.

Sin embargo, se observa una ligera asimetría hacia la izquierda, ya que la cola de la distribución parece extenderse algo más hacia valores negativos que hacia los positivos. Esto sugiere que existen más casos con valores relativamente bajos de *escs_trend*, lo que podría interpretarse como una mayor presencia de observaciones en las que la tendencia del estatus socioeconómico evoluciona de manera menos favorable.

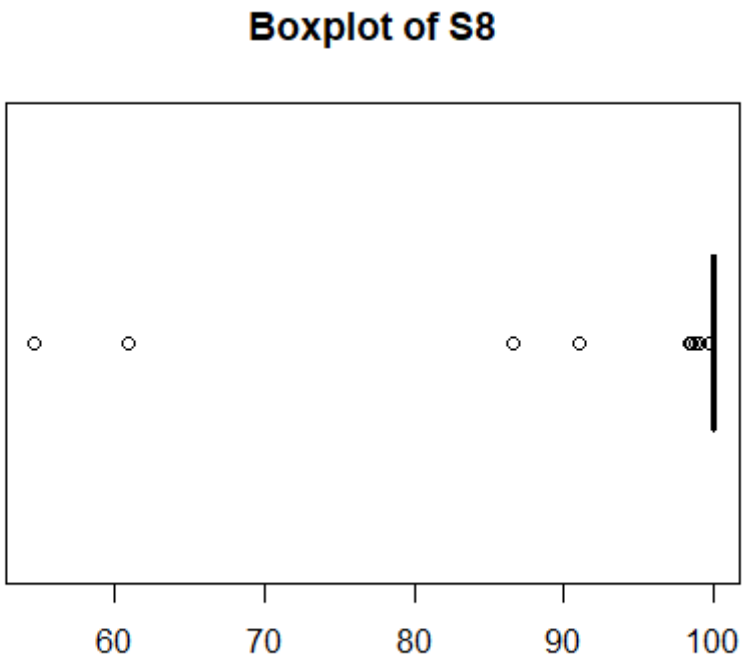


Tabla 3:Boxplot de la variable Proportion of schools with access to the internet for pedagogical purposes [Lower secondary]

El boxplot correspondiente a la variable S8 muestra una distribución altamente concentrada en valores cercanos a 100, lo que indica que en la gran mayoría de las observaciones el acceso a internet con fines pedagógicos en la educación secundaria inferior es prácticamente universal. Esta fuerte concentración en el límite superior de la escala provoca que la caja del boxplot sea extremadamente estrecha, reflejando una variabilidad muy reducida en la parte central de la distribución. En otras palabras, la

mayoría de los países presentan valores muy similares y cercanos al máximo posible del indicador.

No obstante, el gráfico también revela la presencia de varios valores atípicos situados claramente por debajo del grueso de la distribución. Estos puntos, que aparecen aproximadamente en rangos entre 55 y 95, representan países o casos específicos donde el nivel de acceso a internet pedagógico es considerablemente menor que en el resto de la muestra. Este patrón coincide con lo observado previamente en los estadísticos descriptivos, donde se señalaba que, aunque la mayor parte de los valores de los indicadores de infraestructura se situaba en torno a 100, existían algunos valores mínimos sensiblemente más bajos.

Por tanto, en este caso los outliers no corresponden a valores extremadamente altos, sino a observaciones relativamente bajas dentro de una variable que está muy concentrada en su valor máximo. Este fenómeno puede interpretarse como evidencia de desigualdades entre países en el acceso a infraestructuras digitales educativas: mientras que la mayoría ha alcanzado niveles muy elevados de acceso, un pequeño grupo presenta valores significativamente inferiores, lo que genera los puntos atípicos observados en el boxplot.

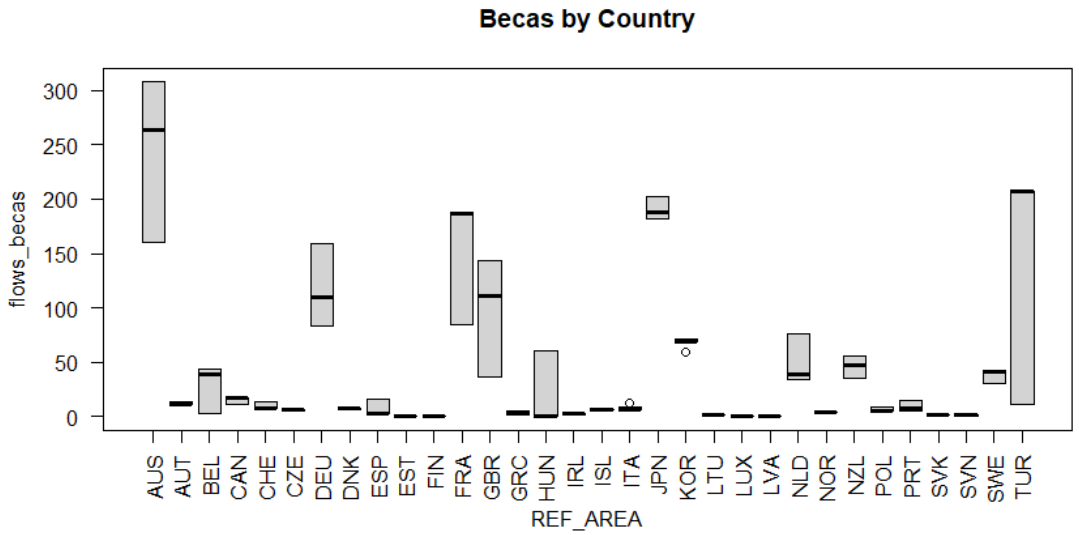


Tabla 4: Boxplot por país de la variable Becas

El boxplot de flows_becas por país pone de manifiesto una alta heterogeneidad entre países en el volumen de flujos de becas. Mientras que algunos países presentan valores muy elevados y con una amplia dispersión, como Australia (AUS), Austria (AUT),

Francia (FRA), Alemania (DEU) o Suecia (SWE), otros mantienen niveles muy reducidos y con escasa variabilidad. Esta diferencia sugiere que las políticas o programas de becas varían considerablemente entre sistemas educativos, tanto en intensidad como en cobertura.

En varios países se observa además un rango intercuartílico amplio, lo que indica que los valores de becas fluctúan de forma notable dentro de cada país a lo largo de años observados. Este fenómeno puede estar relacionado con cambios en políticas educativas, variaciones en el número de beneficiarios o diferencias entre los años considerados en el análisis. En contraste, algunos países presentan cajas muy pequeñas o prácticamente inexistentes, lo que refleja niveles bajos y relativamente estables de flujos de becas.

Asimismo, el gráfico permite identificar la presencia de valores atípicos en determinados países. Estos puntos representan observaciones donde el volumen de becas es inusualmente alto o bajo en comparación con el resto de los valores del mismo país. Este patrón es coherente con lo observado previamente en los estadísticos descriptivos, donde la variable flows_becas mostraba una fuerte dispersión y un valor máximo muy elevado en relación con la mediana. En consecuencia, algunos de estos valores extremos contribuyen a explicar la diferencia entre la media y la mediana detectada anteriormente.

En conjunto, el gráfico confirma que la distribución de los flujos de becas no es homogénea entre países, sino que presenta fuertes diferencias estructurales. Mientras que algunos sistemas educativos muestran niveles elevados y variables de financiación mediante becas, otros presentan valores mucho más reducidos, lo que puede reflejar distintos enfoques en las políticas de apoyo económico al alumnado.

Se ha preferido no mostrar ninguna gráfica de boxplots por año, puesto que su interés para el desarrollo del trabajo es limitado, al no aportar información relevante.

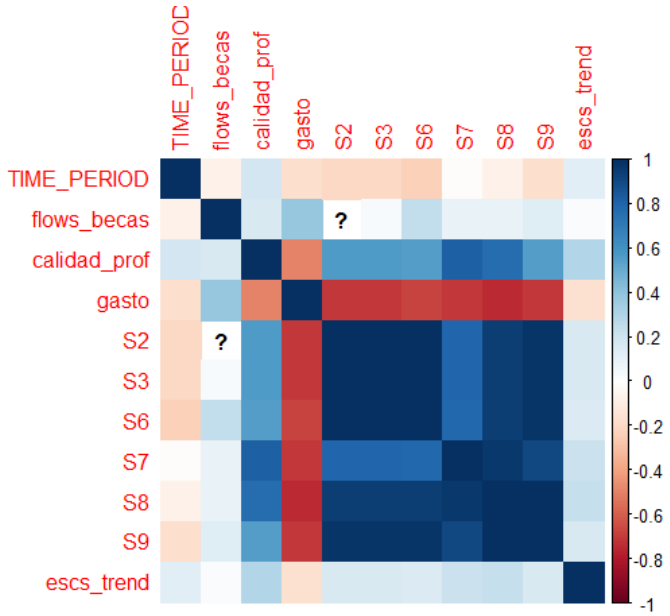


Tabla 5: Matriz de correlación de las variables de entrada

La matriz de correlación muestra varios patrones relevantes en la relación entre las variables del dataset. En primer lugar, destaca la existencia de un bloque muy fuerte de correlaciones positivas entre los indicadores de infraestructura escolar (S2, S3, S6, S7, S8 y S9). Estas variables presentan coeficientes cercanos a 1, lo que indica que están altamente relacionadas entre sí. Este resultado es coherente desde el punto de vista conceptual, ya que todos estos indicadores capturan distintos aspectos del acceso a recursos básicos en los centros educativos.

Esta fuerte correlación de las variables de infraestructura parece crearnos un grave problema de multicolinealidad que puede afectar a la viabilidad modelo, una vez empiece la programación. De este modo, para evitar generar un ruido innecesario que pueda perjudicar los resultados de la investigación, se decidió unir todos los índices de infraestructura en uno único, obteniendo el siguiente mapa de correlación corregido:

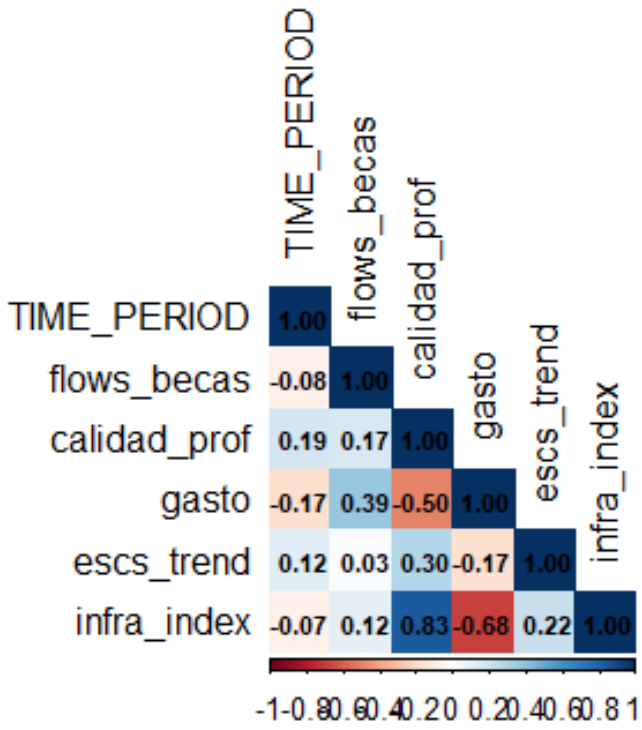


Tabla 6: Mapa de correlación de las variables de entrada corregido para evitar la multicolinealidad

Una vez corregido, vemos como la variable gasto presenta correlaciones negativas relativamente intensas con los indicadores de infraestructura, lo cual resulta llamativo. Esto podría indicar que en aquellos contextos donde el gasto es más elevado, los niveles de acceso a ciertas infraestructuras no necesariamente son mayores, o bien que podrían existir diferencias estructurales en cómo se mide o distribuye el gasto entre países. No obstante, este resultado también podría estar influido por la presencia de outliers previamente identificados en esta variable, especialmente por valores extremadamente altos que distorsionan las relaciones lineales.

Por otro lado, las variables flows_becas y calidad_prof muestran correlaciones positivas moderadas con los indicadores de infraestructura, lo que sugiere que mejores condiciones en los centros educativos tienden a asociarse con mayores niveles de calidad del profesorado y, en menor medida, con mayores flujos de becas. Sin embargo, estas relaciones no son especialmente fuertes, lo que indica que existen otros factores que también influyen de manera significativa en estas variables.

En cuanto a escs_trend, se observa que sus correlaciones con el resto de variables son relativamente débiles, lo que sugiere que la evolución del estatus socioeconómico del alumnado no está fuertemente asociada de forma lineal con el resto de los indicadores

considerados. Este resultado es consistente con el análisis previo de su distribución, donde se apreciaba una variabilidad moderada y ausencia de patrones extremos.

Tratamiento de los datos de salida

En cuanto a las variables de salida, toda la información se recopiló de distintas bases de datos de *Eurostat*. Para crear las variables deseadas, se llevaron a cabo los siguientes pasos:

Para analizar la inserción laboral en los distintos países, se utilizó el dataset Employement rates by educational atainment level (lfsa_ergaed). Para llegar a la información que deseábamos, hubo que hacer cierta limpieza de datos. En primer lugar, se eligió la horquilla de edad de 20 a 64 años, al poder considerar este intervalo de años como los de actividad laboral, acto seguido consideramos que no tenía sentido hacer distinción por géneros, por lo cual se eligió el género como *Total* y, por último, solo se tuvieron en cuenta los datos cuyo índice ISCED11, relativo al nivel educativo alcanzado, fuese de educación superior (ED5-8).

Para analizar la correspondencia entre educación y empleo, se utilizó el dataset Over-qualification rates by citizenship (lfsa_eoqgan). Al igual que en la variable de inserción, se utilizó el índice de edad de 20 a 64 años y el género *Total.*

En cuanto a la brecha salarial de género, se utilizó el dataset Median hourly earnings, all employees (excluding apprentices) by sex (earn_ses_pub2s). Para extraer la información que necesitábamos de esta base de datos, se filtraron los datos por sexo, para luego crear una columna adicional en la que figurase la resta entre el salario por hora entre hombres y mujeres. Esta diferencia será la que marcará la brecha.

Para el rendimiento, se busca que los puestos de mayor responsabilidad estén ocupados por personas con mayor grado de estudios. Para esto, se utilizó el dataset Employees by educational attainment level and occupation (edat_lfs_9905). Para el grado de estudios, al igual que en el dataset de inserción laboral, se utilizó el ISCED11 5-8. Para el nivel de responsabilidad en el puesto de trabajo, por su parte, se filtró por el código ISCO08 OC1, haciendo referencia al mayor grado jerárquico dentro de una empresa. Para adecuarse al resto de datasets, se tuvo en cuenta también solo el género *Total*. La única diferencia con

respecto a las bases de datos anteriores es la necesidad de utilizar el intervalo de edad de 15 a 64 años, al no figurar en los datos de origen una horquilla de 20 a 64, como en los otros.

Por último, para analizar la movilidad social, se tuvo en cuenta el dataset Population in private households by educational attainment level, country of birth and degree of urbanisation (edat_lfs_9915). Para alcanzar los datos deseados, se filtró por el intervalo de edad de 15 a 64 años, género *Total* por los mismos motivos que en el rendimiento; por grado de educación ISCED11 5-8 y por grado de urbanización DEG3, es decir, población rural.

Una vez delimitadas todas las variables de salida, se procedió a unirlas en un nuevo dataset final, utilizando como clave de unión el país y el año, para que de este modo coincidiese la estructura con el dataset de entrada. Para seguir haciéndolos plenamente compatibles, decidimos filtrar los años del dataset de salida para que fuesen iguales a los de entrada, es decir, utilizando los años del informe PISA: 2012, 2015 y 2018.

Análisis exploratorio de los datos de salida

El dataset de salida presenta una estructura sólida, compuesta por 105 observaciones y 7 variables, que cubren el periodo PISA de 2012, 2015 y 2018 y un total de 36 países pertenecientes a la lista de encuestados por EUROSTAT, la cual incluye tanto a países de la Unión Europea y Zona Euro como a países con estrechas relaciones económicas y políticas con éstos, véase Noruega, Reino Unido y Turquía. La comprobación mediante la agrupación por país y año arroja cero registros repetidos, lo que garantiza que cada unidad de análisis es única y está lista para el modelado de series temporales.

En cuanto al comportamiento de las variables, la inserción laboral tiene una media del 82.49%, aunque la brecha entre el mínimo de 66.60% y el máximo de 92.20% sugiere una disparidad significativa entre los sistemas educativos europeos. Por otro lado, la variable de sobre cualificación presenta una media de 27.08%, llegando a alcanzar picos preocupantes de más del 50% en ciertos países. Esto indica un desajuste laboral donde, a pesar de los buenos índices de inserción, una parte considerable de los graduados desempeña funciones por debajo de su nivel formativo.

Un aspecto metodológico observado en el resumen estadístico es la existencia de valores vacíos, con una incidencia especialmente elevada en la variable de brecha salarial, donde se registran 71 ausencias sobre un total de 106 observaciones. Esta situación puede estar relacionada con la irregularidad en la recopilación de datos salariales por parte de las fuentes oficiales en determinados años, al hacerse la encuesta de salarios en función de sexo de forma cuatrianual, no coincidiendo, de este modo, a la perfección con los años de los datos PISA.

Finalmente, la movilidad rural, con un rango que oscila entre el 3.80% y el 46.84%, destaca como el indicador con mayor variabilidad relativa. Esto sugiere que el éxito del sistema educativo no solo debe medirse en términos de empleo bruto, sino en la capacidad de los graduados para integrarse en diversos entornos geográficos, evitando la polarización del talento en los núcleos urbanos y segregando las poblaciones rurales a un estancamiento evolutivo en materia de educación.

Después de llevar a cabo estos pasos, se procedió a realizar una serie de gráficas que explicasen la distribución de cada una de las variables. Se realizaron: histogramas, boxplots, boxplots por país, boxplots por año y una matriz de correlación. A continuación, se hará un análisis de las gráficas más destacables:

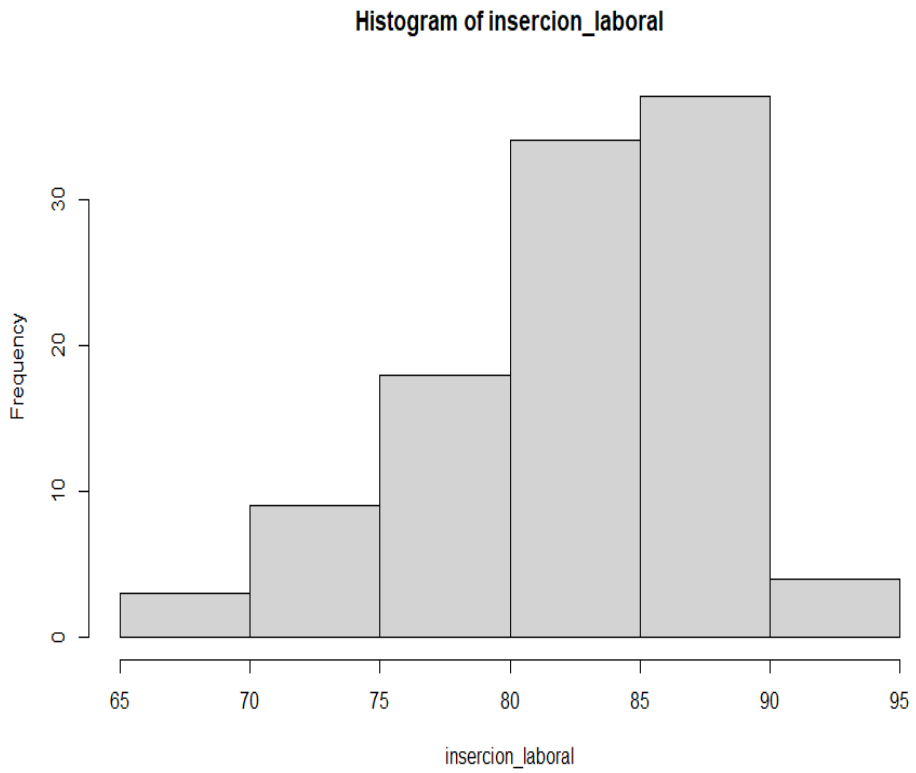


Tabla 7: Histograma de inserción Laboral

Como se constata en la gráfica, se aprecia que la mayoría de las observaciones se concentran en el rango de valores altos, específicamente entre el 80% y el 90%, lo que indica que, en general, los niveles de inserción en la muestra son elevados. Esto establece un estándar de desempeño alto para la mayoría de los registros analizados, algo esperable en el contexto social y económico de la Unión Europea.

Se observa una marcada asimetría hacia la izquierda, lo que parece indicar un sesgo negativo. Mientras que la cola derecha de la distribución es corta y se detiene antes del 95%, la cola izquierda se extiende de manera mucho más prolongada hacia valores inferiores al 70%. Esto sugiere que existe una mayor presencia de casos con resultados desfavorables en comparación con los casos de éxito excepcional, marcada seguramente por la inclusión de países aún en vías de desarrollo económico, véase países de Europa del Este.

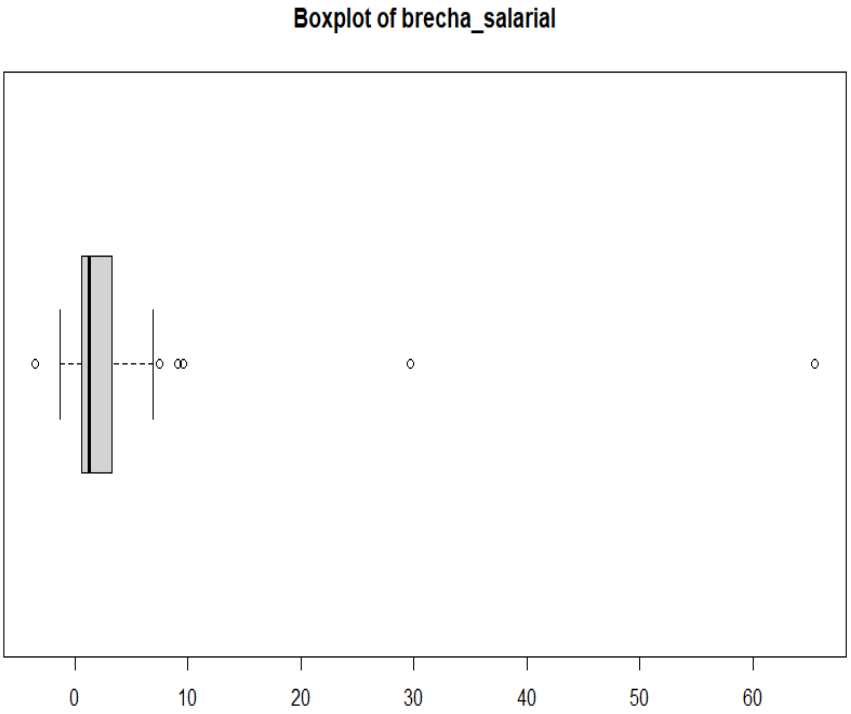


Tabla 8: Boxplot de la brecha salarial

La forma del boxplot de brecha salarial es notablemente compacta. Esta concentración sugiere una alta homogeneidad en la mayoría de los países analizados, donde los resultados de equidad salarial tienden a ser estables y positivos, es decir, cercanos a la paridad.

Asimismo, aunque la mayor parte de los datos se sitúa en el rango inferior a 10, se identifican varias observaciones extremas a lo largo del eje. Destacan especialmente dos puntos situados en niveles críticos (aproximadamente en 30 y 65). Estos casos pueden considerarse valores atípicos severos que distorsionan la media del conjunto.

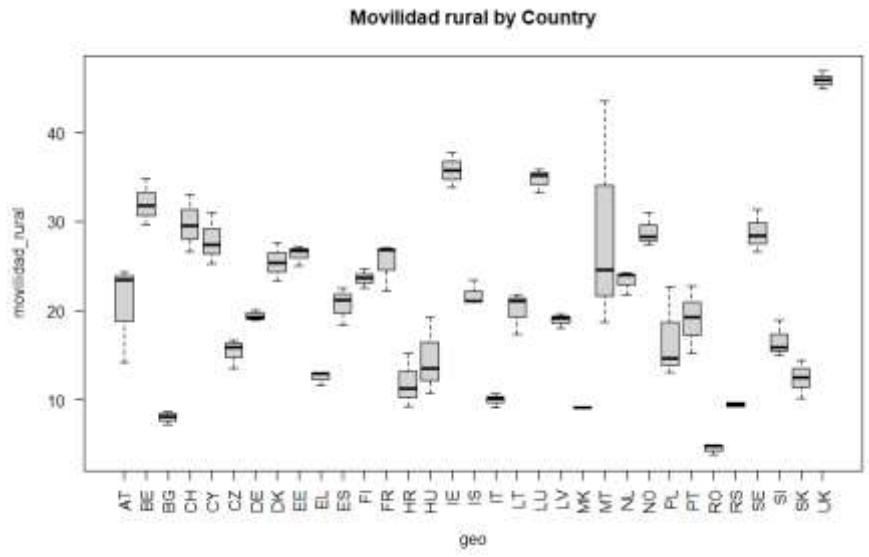


Tabla 9: Boxplot por país de la movilidad rural

Se aprecia que la mayoría de las observaciones por país se concentran en rangos muy distintos unos de otros, lo que indica que la movilidad rural es un factor fuertemente dependiente del contexto nacional. Mientras que países como Reino Unido presentan los niveles más altos de la muestra, superando el 45%, otros como Rumanía o Bulgaria se sitúan de forma estable en niveles inferiores al 10%. Esta brecha evidencia realidades territoriales y dinámicas de mercado laboral muy diversas dentro de la Unión Europea y a priori relacionadas con el desarrollo económico de las distintas naciones.

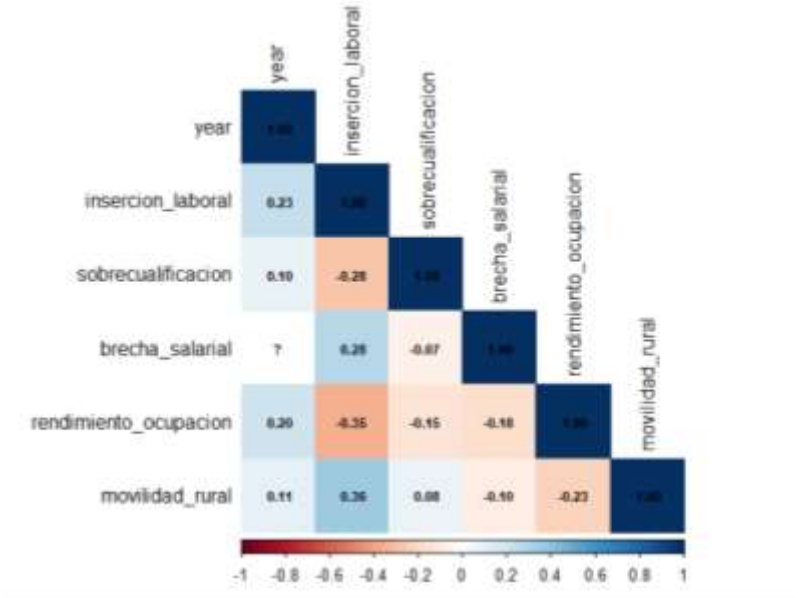


Tabla 10: Mapa de correlación de las variables de salida

La observación más relevante es la correlación negativa moderada (-0.35) entre el rendimiento_ocupación y la insercion_laboral, lo que sugiere un fenómeno de permuta estructural. En algunos contextos, una expansión rápida de la cantidad de empleo podría estar asociada a una disminución en la calidad o el encaje técnico de dichos puestos. Asimismo, la insercion_laboral muestra una relación positiva con la movilidad_rural (0.36), indicando que los mercados laborales más dinámicos y capaces de absorber graduados tienden a ser aquellos que también fomentan una distribución territorial más equilibrada del talento.

DETALLE TÉCNICO DEL PROYECTO

Tratamiento de los datos para la generación del modelo

Para definir el marco geográfico y político del estudio, la muestra se ha configurado seleccionando un conjunto heterogéneo de naciones que comparten estándares de reporte internacionales, permitiendo una comparativa entre diferentes modelos de gobernanza. Principalmente, el trabajo se centra en países miembros de la Unión Europea (UE), incluyendo tanto a las potencias de la Zona Euro (por ejemplo: Alemania, Francia, España e Italia) como a economías del este y centro de Europa (véase Polonia, República Checa o Rumanía). Asimismo, la muestra trasciende las fronteras de la Unión al integrar países miembros de la OCDE que no pertenecen al bloque comunitario, como es el caso de Reino Unido, Noruega y Suiza, lo que permite contrastar la eficacia de los sistemas educativos europeos bajo distintos marcos institucionales y monetarios.

Una vez finalizado el proceso de creación de la base de datos y explorado su contenido, procedemos a la fase de modelaje. Para esto, decidimos cambiar el lenguaje de programación. Si bien el tratamiento primario de los datos y la confección del dataset final se hizo en lenguaje Rstudio, el modelaje decidimos hacerlo en Python, debido a razones de comodidad, organización visual y maestría. Para esto, exportamos en formato .csv el dataset final del guion de R, para cargarlo al cuaderno de Python mediante la librería Pandas.

Una vez cargada la base de datos de variables de entrada, hicimos un filtrado de países para que figurasen solo aquellos del dataset de salida. Si bien el dataset contaba con información de todos los países que forman parte de la OCDE, solo aquellos que formen parte de la lista de Eurostat podrán compararse con los resultados de las variables de salida. Por tanto, consideramos que utilizar los datos de todos los países solo contribuiría con más ruido al modelo sin aportar realmente resultados interesantes para responder a la cuestión central de la investigación.

En una segunda instancia, decidimos crear un índice de infraestructuras haciendo la media de todas las variables que la conformaban y que habíamos bautizado como SX (siendo X el número que figura en la tabla 1). Esto se hizo para poder crear una base de datos que permitiese hacer un estudio primario no exhaustivamente avanzado, pero que sí nos

permitiese hacernos una idea de lo que nos íbamos a encontrar de ahí en adelante. Para este estudio, decidimos implementar un clustering jerárquico. La creación del índice de infraestructura fue muy útil, puesto que, sin ella, la matriz no tenía un tamaño que permitiese analizar los datos con dicho procedimiento.

Clustering jerárquico

Antes de proceder al clustering jerárquico, hubo que solucionar el problema de los valores vacíos del dataset, siendo *conditio sine qua non* para poder hacer la clasificación que no exista ninguno de estos valores en la base de datos. Para realizar esto, se decidió hacer una interpolación por país, de modo que aquellos valores ausentes serían rellenados con la tendencia del mismo país. En caso de que la observación del país fuese completamente inexistente, es decir, no hubiese ningún registro para ninguno de los años en ese país; se rellenaría con la media del resto de países.

Una vez rellenados los valores vacíos, procedimos al clustering haciendo las medias históricas de cada país y utilizando el método Ward con distancia euclídea, al ser este el de uso más común en estos procesos, obteniendo el siguiente resultado:

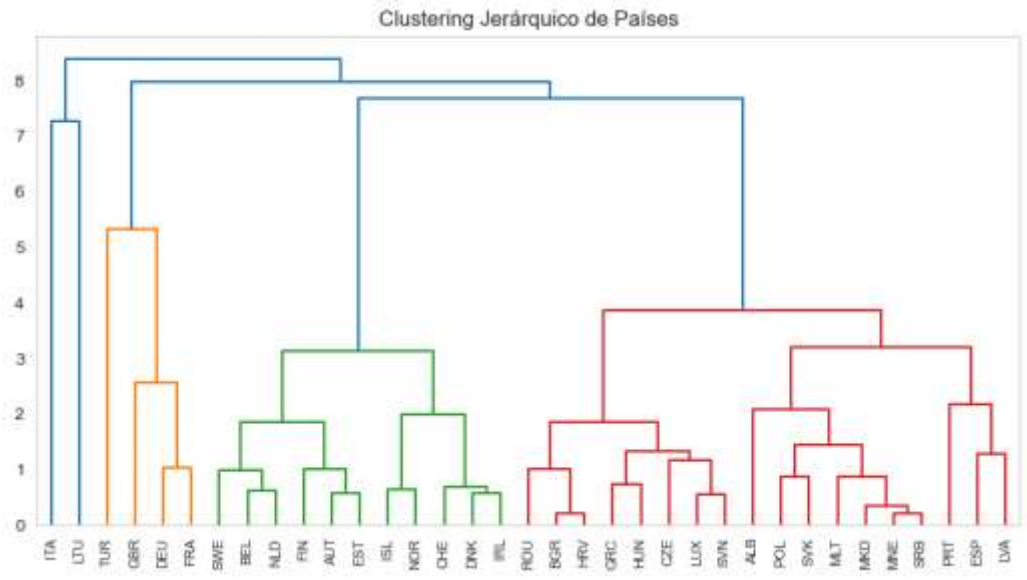


Tabla 11: Clustering jerárquico de países en función de las variables de educación de calidad (entrada)

En el gráfico podemos observar cuatro grupos claramente diferenciados. En primer lugar, se observa una divergencia clara en las tendencias de Italia (ITA) y Lituania (LTU) con

valores muy extremos que pueden llegar a ser atípicos y que no llegan a identificarse en ninguno de los otros clústeres.

El clúster verde muestra distancias muy bajas, lo que parece indicar perfiles muy similares, mientras que el clúster rojo parece más heterogéneo. Aun así, se pueden ver patrones significativos en ambos, como la estrecha relación entre España y Portugal o la inclusión de varios de los países nórdicos en el clúster verde, lo cual parece implicar que existe un factor cultural muy fuerte en la clasificación de los distintos grupos.

Esto puede seguir observándose en el clúster naranja, donde se agrupan las tres grandes potencias económicas europeas (Reino Unido, Francia y Alemania).

Sistema de scores

Una vez visualizados de forma preliminar los diferentes grupos en base a nuestra aproximación a la calidad educativa, decidimos hacer un sistema de puntos para clasificar a los países. La clasificación responde a los siguientes criterios:

- El flujo de becas deberá situarse por encima de la media: este criterio se incluye porque un mayor flujo de becas puede interpretarse como un indicador de apoyo institucional al acceso equitativo a la educación. Aquellos países que superen la media en este aspecto estarían, en principio, facilitando que estudiantes de distintos contextos socioeconómicos puedan continuar su formación, reduciendo así barreras de acceso.
- El gasto en educación deberá situarse por encima de la media: el gasto educativo se considera un proxy del compromiso del país con su sistema educativo. Situarse por encima de la media implicaría una mayor inversión en recursos, personal e infraestructuras, lo que, en términos generales, podría traducirse en mejores condiciones de aprendizaje y enseñanza.
- La tendencia escs deberá situarse por encima de la media: la tendencia del índice escs se incorpora para captar la relación entre el contexto socioeconómico del alumnado y su rendimiento. Un valor por encima de la media podría sugerir un entorno más favorable o una menor desigualdad en las condiciones de partida, lo que facilitaría el desarrollo académico.

- La calidad del profesorado y el índice de infraestructura penalizarán en caso de encontrarse por debajo de la media: la decisión de que la calidad del profesorado y el índice de infraestructura penalicen responde a lo que se comentó en el análisis exploratorio de los datos sobre el valor de los outliers. Al situarse la media y la mediana cercana a 100, aquellos países que se alejen de esta cifra se encontrarán muy por detrás del resto de países en cuanto a lo que nosotros consideramos una educación de calidad.

Antes de asignar las puntuaciones, no obstante, se decidió normalizar las variables de gasto y flujo de becas, con el objetivo de que los datos atípicos no desajustasen en sobremedida el modelo. Una vez asignados los puntos, también se calculó una columna de mejoría a lo largo de los años, obteniendo el siguiente resultado de los cinco países con mayor mejoría a lo largo de los años:

País	Año	Flujo de becas	Calidad prof	Gasto	Infra index	Escs trend	Score Desempeño	Mejora Relativa
FRA	2015	187.75	98.61	8.27	100.00	-0.087	117.84	**50.19**
TUR	2015	109.30	98.61	12.89	100.00	-1.560	129.31	**49.71**
HUN	2018	60.23	98.61	7.36	100.00	-0.121	43.36	**39.26**
GBR	2015	110.42	98.61	12.23	99.75	0.072	137.09	**36.89**
TUR	2018	207.30	98.61	11.61	100.00	-1.245	162.00	**32.69**

Tabla 12: Países con mayores mejoras relativas en calidad educativa

El análisis del ranking de Mejora Relativa revela que países como Francia (FRA) y Turquía (TUR) lideran la eficiencia del modelo, logrando puntuaciones de desempeño muy por encima de lo esperado según su nivel de inversión. Destaca especialmente el caso de Turquía, que aparece en dos periodos distintos (2015 y 2018), lo cual sugiere una trayectoria de mejora sostenida. Al observar las variables subyacentes, se aprecia que este éxito parece no responder únicamente al gasto bruto, sino a una gestión optimizada de los recursos: mientras que Francia mantiene un gasto moderado (8,27), compensa su posición con un flujo de becas muy elevado (187,75). Por otro lado, Turquía, a pesar de enfrentarse a un contexto socioeconómico más complejo (indicado por el valor negativo de -1.560 en escs_trend), logra compensar esta desventaja competitiva mediante la movilización máxima de ayudas al alumnado, alcanzando el techo del modelo en flujos normalizados.

Una vez hecha la clasificación, procedimos a realizar un modelo de Random Forest que nos explicase qué variables son más importantes a la hora de hacer la clasificación. El resultado del modelo nos informaría sobre qué variables son más significativas del acercamiento que hemos hecho a la educación de calidad.

Para abordar la validez metodológica del estudio, es necesario precisar que, si bien el Random Forest y el sistema de Scoring desarrollado no pretenden erigirse como modelos de robustez estadística absoluta, su aplicación resulta suficientemente útil para los objetivos de este trabajo. Estos algoritmos se han seleccionado por su capacidad para manejar relaciones no lineales y capturar la importancia de variables en conjuntos de datos complejos. Así, estas herramientas actúan como marcos analíticos que facilitan la identificación de tendencias, la clasificación de patrones nacionales y la generación de una jerarquía de factores clave, cumpliendo con éxito la función de proporcionar una base técnica sólida para la toma de decisiones y la discusión académica dentro del alcance de la investigación. Estos sistemas se plantean como herramientas de carácter exploratorio, útiles para orientar el análisis y detectar posibles patrones, pero no como instrumentos que permitan extraer conclusiones definitivas.

```
from sklearn.ensemble import RandomForestClassifier
import seaborn as sns

X_rf = df_country[cols_estudio]

rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_rf, y)

importancias = pd.DataFrame({
    'Variable': cols_estudio,
    'Importancia': rf_model.feature_importances_
}).sort_values(by='Importancia', ascending=False)
```

Tabla 13: Código para la implementación del Random Forest

El uso de estos parámetros específicos en el entrenamiento del Random Forest responde a la necesidad de equilibrar la precisión del modelo con la solidez metodológica. Al configurar n_estimators=100, estamos creando un "bosque" de cien árboles de decisión independientes que votan para determinar el resultado final; esta cantidad es un estándar que permite reducir drásticamente el riesgo de sobreajuste y suavizar el error del modelo sin disparar el coste computacional. Por otro lado, la inclusión de random_state=42 se

utilizó para fijar la semilla del generador aleatorio. Dado que el Random Forest selecciona variables y muestras de forma aleatoria en cada iteración, sin esta semilla los resultados variarían cada vez que ejecutamos el código; al fijarla, garantizamos la replicabilidad de los hallazgos.

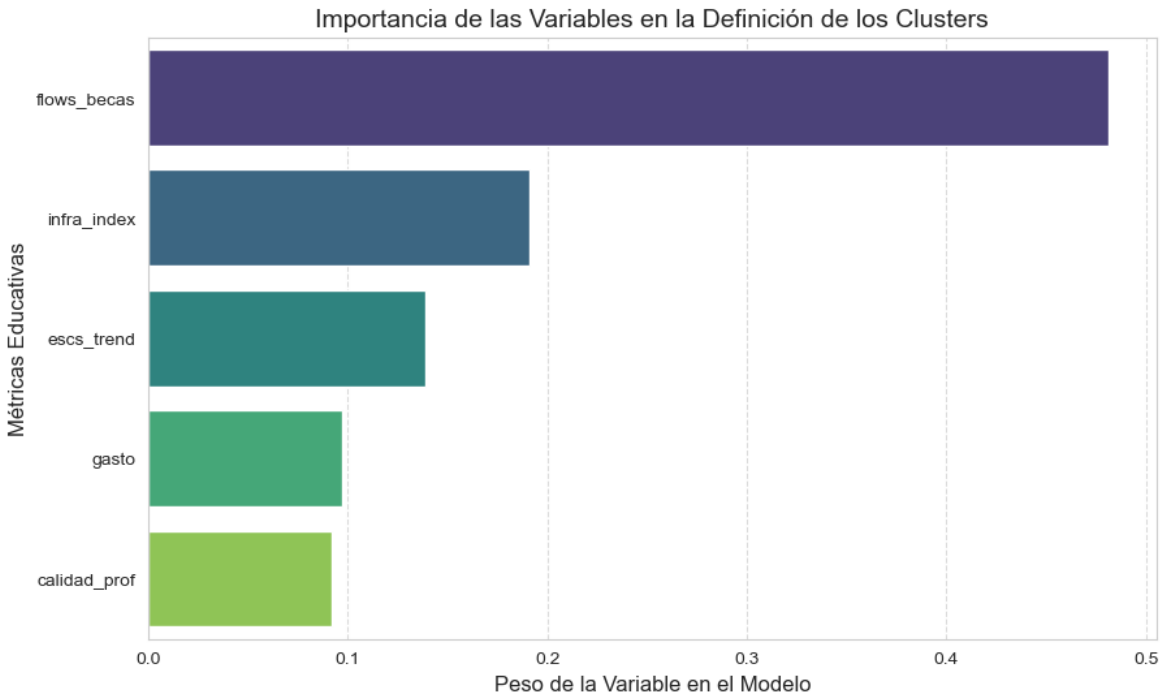


Tabla 14: Importancia de las variables en la definición de clústeres

El resultado de la implementación del Random Forest desemboca en este gráfico de barras sobre la importancia de las variables. De acuerdo con los resultados del modelo, la variable flows_becas parece emerger como el predictor más influyente, con un 45%, lo que sugiere que la movilidad financiera de los estudiantes y el apoyo directo a través de becas son los factores que más discriminan entre un sistema educativo de alto rendimiento y uno rezagado. A gran distancia, le siguen el índice de infraestructura y el contexto socioeconómico, mientras que el gasto bruto y la calidad percibida del profesorado muestran pesos significativamente menores. Estos datos parecen proponer que, para mejorar la eficiencia del sistema, la gestión estratégica de las becas tiene un impacto más decisivo en la estructura del modelo que el simple aumento del gasto educativo general.

Para seguir ahondando en las diferencias en el nivel educativo de los países, quisimos hacer una gráfica donde se comparase la evolución de algunos países en función de la evolución de sus scores:

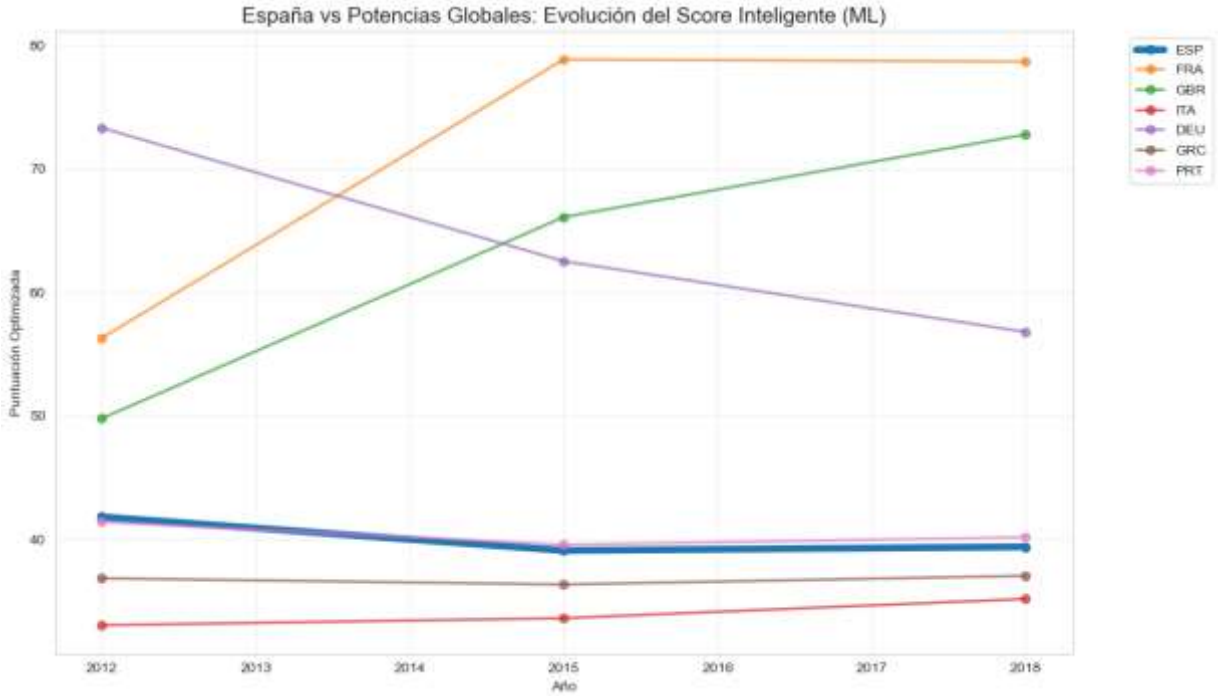


Tabla 15: Comparación de la puntuación de calidad educativa de España, Francia, Reino Unido, Italia, Alemania, Grecia y Portugal

La gráfica sugiere una tendencia de países del centro de Europa a obtener una calificación más alta, mientras que países del suroeste se agrupan en una zona de puntuaciones más bajas. Esto viene a reforzar la narrativa que observábamos previamente en el clustering jerárquico y cómo la cultura y el desarrollo económico de las diferentes zonas geográficas parece jugar un papel determinante en la clasificación de los países.

A modo de resumen del proceso de clasificación de los países en función de las variables que consideramos pueden acercarse a lo que venimos definiendo en este trabajo como educación de calidad, procedimos a crear un mapa de calor de todos los países del estudio donde quedase patente su clasificación a lo largo de los años PISA (2012, 2015 y 2018):

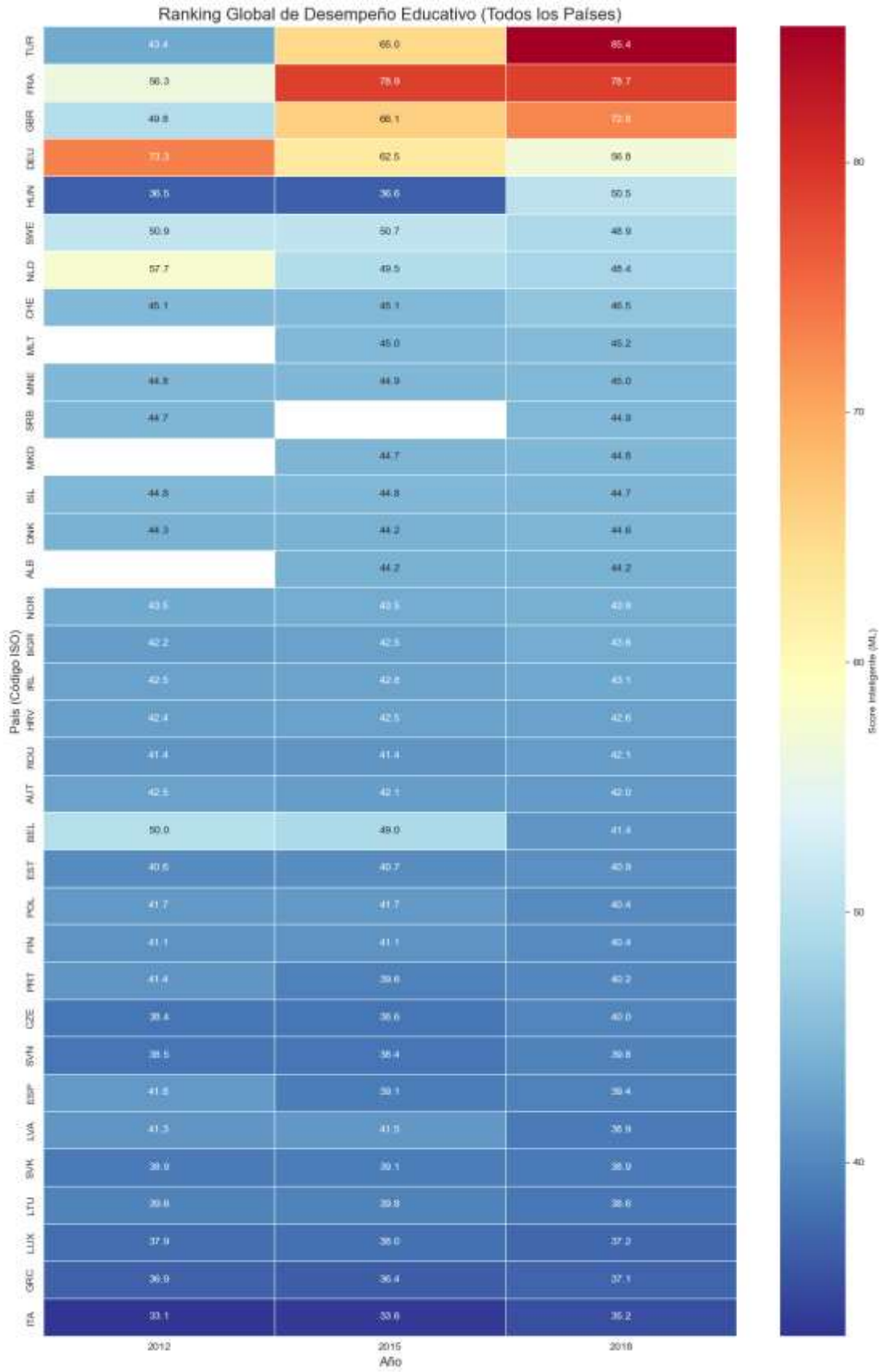


Tabla 16: Mapa de calor de los scores educativos

Habiendo finalizado el estudio clasificatorio de las variables de entrada, exportamos un dataset final que contuviese las puntuaciones, para poder empezar a relacionar estos resultados con los de las variables de salida.

Inserción de las variables de salida

Una vez analizadas las variables de entrada, procedemos a cargar en Python el dataset con las variables de salida. Para poder hacer una comparación entre ambos, tuvimos que hacer previamente ciertos ajustes. En primer lugar, cambiar el nombre de las variables *geo* y *year* del dataset de salida por *REF_AREA* y *TIME_PERIOD*, respectivamente, para poder unirlos. En segundo lugar, tuvimos que cambiar la nomenclatura de los códigos de país, al figurar los datos de entrada en un sistema de tres letras y los de salida en un sistema de dos letras. De este modo, un país como España que figura en el dataset de salida como ES, pasaría a ser ESP, siendo este el nombre con el que se le reconoce en el sistema de clasificación ISO de la OCDE. Una vez hechos estos ajustes, se procedió a crear un dataset conjunto de entrada y salida, usando como clave de unión el país y el año (REF_AREA y TIME_PERIOD).

```
dataset_salida = pd.read_csv("C:/TFG/Analytics/Documentación/BBDD/dataset_salida_PISA_years.csv")


if 'geo' in dataset_salida.columns:
    dataset_salida = dataset_salida.rename(columns={'geo': 'REF_AREA', 'year': 'TIME_PERIOD'})


if dataset_salida['REF_AREA'].str.len().max() == 2:
    mapa_iso = {
    'AT': 'AUT', 'BE': 'BEL', 'BG': 'BGR', 'CY': 'CYP', 'CZ': 'CZE',
    'DE': 'DEU', 'DK': 'DNK', 'EE': 'EST', 'EL': 'GRC', 'ES': 'ESP',
    'FI': 'FIN', 'FR': 'FRA', 'HR': 'HRV', 'HU': 'HUN', 'IE': 'IRL',
    'IT': 'ITA', 'LT': 'LTU', 'LU': 'LUX', 'LV': 'LVA', 'MT': 'MLT',
    'NL': 'NLD', 'PL': 'POL', 'PT': 'PRT', 'RO': 'ROU', 'SE': 'SWE',
    'SI': 'SVN', 'SK': 'SVK', 'UK': 'GBR', 'NO': 'NOR', 'CH': 'CHE'
}

    dataset_salida['REF_AREA'] = dataset_salida['REF_AREA'].map(mapa_iso).fillna(dataset_salida['REF_AREA'])


df_master = pd.merge(df_clean, dataset_salida, on=['REF_AREA', 'TIME_PERIOD'], how='inner')
```

Tabla 17: Carga del dataset de salida y adecuación de nombres de paises al sistema ISO de la OCDE

Acto seguido, procedimos a la creación de scores de las variables de salida, dónde:

- Buscamos inserción laboral por encima de la media: este indicador se utiliza para evaluar la capacidad del sistema educativo de facilitar la transición al mercado de trabajo. Un nivel de inserción laboral superior a la media sugeriría que la

formación recibida está alineada, al menos en parte, con las demandas del mercado, favoreciendo la empleabilidad de los individuos.

- Buscamos rendimiento por ocupación por encima de la media: el rendimiento por ocupación permite analizar hasta qué punto los niveles educativos se traducen en posiciones laborales acordes. Valores por encima de la media podrían indicar una mejor correspondencia entre formación y empleo desempeñado, reflejando una mayor eficiencia del sistema educativo en la asignación de talento.
- Buscamos movilidad rural por encima de la media: este criterio se incluye para captar la capacidad del sistema de generar oportunidades independientemente del origen geográfico. Una mayor movilidad rural podría interpretarse como una reducción de las desigualdades territoriales, permitiendo que individuos de zonas menos favorecidas accedan a mejores oportunidades educativas y laborales.
- Buscamos sobre cualificación lo más baja posible: la sobre cualificación se considera un desajuste entre el nivel educativo alcanzado y el requerido por el puesto de trabajo. Niveles bajos indicarían una mejor adecuación entre formación y empleo, evitando la infrautilización del capital humano.
- Buscamos una brecha salarial lo más baja posible: La brecha salarial se incorpora como medida de desigualdad en los resultados del mercado laboral. Valores reducidos podrían reflejar una distribución más equitativa de los ingresos, lo que, en conjunto con el resto de indicadores, contribuiría a una evaluación más favorable del sistema en términos de equidad y eficiencia.

Obteniendo los siguientes resultados:

País	Año	Score Salida
LUX	2018	**88.11**
LUX	2012	**81.89**
LTU	2018	**77.12**
LUX	2015	**76.85**
CHE	2018	**73.18**
SVN	2018	**72.76**
MLT	2018	**72.35**
BEL	2018	**72.17**
FIN	2018	**70.52**
NLD	2018	**70.00**

Tabla 18: 10 países con mejores scores de salida

Luxemburgo no solo lidera el ranking, sino que ocupa tres de las cuatro primeras posiciones en diferentes periodos temporales (2012, 2015 y 2018). Su puntuación máxima de 88,11 en 2018 marca el techo de rendimiento de toda la muestra. Esto sugiere que Luxemburgo posee un sistema altamente eficiente y capaz de traducir sus recursos en resultados laborales excepcionales de forma consistente, distanciándose por más de 10 puntos del grueso de la tabla.

Es destacable la tercera posición de Lituania en 2018 con un score de 77,12. En el contexto del estudio, esto indica que, a pesar de no ser una de las grandes potencias económicas tradicionales de la UE, ha logrado optimizar sus métricas de salida (probablemente impulsado por una alta inserción laboral y baja brecha salarial), superando a países con mayor gasto educativo como Suiza o Finlandia.

Resulta analíticamente relevante que 8 de los 10 mejores registros pertenecen al año 2018. Esto podría interpretarse como una tendencia generalizada de mejora en la eficiencia de los resultados de los sistemas educativos europeos hacia el final de la década, o bien como una recuperación sólida de los mercados laborales tras los efectos de crisis anteriores.

Resultados finales

Una vez creadas las clasificaciones de entrada y salida, procedimos a hacer un gráfico de cuatro ejes donde estuviesen cruzadas ambas. La posición de cada país en esos ejes determinaría su posible grado de eficacia meritocrática. De este modo, procedimos a dividir la gráfica en cuatro sectores:

- El cuadrante de alto score educativo y alto score meritocrático recibe el nombre de EFICIENTES
- El cuadrante de alto score educativo y bajo score meritocrático recibe el nombre de INEFICIENTES
- El cuadrante de bajo score educativo y alto score meritocrático recibe el nombre de ALTO RETORNO
- El cuadrante de bajo score educativo y bajo score meritocrático recibe el nombre de BAJO RETORNO

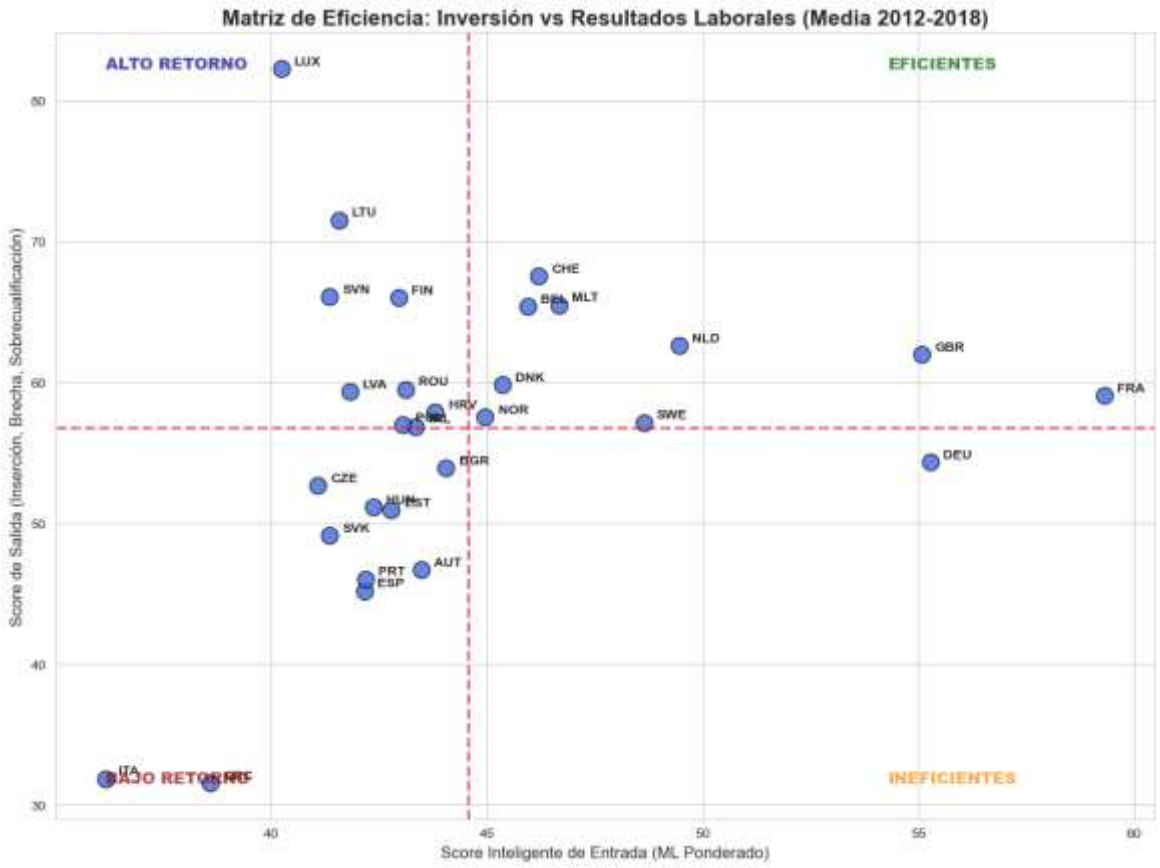


Tabla 19: Matriz de eficiencia meritocrática

En el cuadrante superior izquierdo se sitúan los países de alto retorno, caracterizados por obtener resultados sobresalientes mediante una inversión inteligente y estratégicamente dirigida, sin necesidad de movilizar los volúmenes de gasto más altos de la muestra. Luxemburgo parece erigirse como el líder de este grupo, registrando el mayor score de salida y demostrando una eficiencia excepcional en la garantía de inserción laboral y equidad. En una línea similar, Lituania (LTU) y Eslovenia (SVN) representan casos de estudio de gran interés para la investigación; a pesar de contar con una inversión significativamente menor que la de grandes potencias europeas, logran resultados de salida superiores, posicionándose como referentes de la máxima eficiencia relativa.

El cuadrante superior derecho agrupa a los países más eficientes, naciones con un score de entrada elevado que logran mantener resultados proporcionalmente altos. Dentro de este bloque, Suiza (CHE), Países Bajos (NLD) y Bélgica (BEL) parecen mostrar un equilibrio saludable, donde el alto gasto educativo parece ser absorbido eficazmente por sus mercados laborales. Por el contrario, potencias como el Reino Unido (GBR) y Francia (FRA), aunque se mantienen en el rango de eficiencia, sugieren una dependencia crítica

del volumen de recursos. Francia, en particular, presenta el score de entrada más alto de toda la muestra, lo que sugiere un sistema robusto, pero con un mantenimiento financiero sumamente costoso para sostener sus indicadores de salida.

En el cuadrante inferior izquierdo parecen agruparse los países de bajo retorno, donde tanto la inversión como los resultados son inferiores en comparación con sus homólogos europeos. España (ESP) y Portugal (PRT) muestran posiciones casi idénticas, enfrentando el doble reto de mejorar la calidad de su inversión de entrada y reformar las dinámicas de salida para reducir problemas como la sobre cualificación. No obstante, los casos más críticos son los de Italia (ITA) y Grecia (GRC); situados de forma aislada en el extremo inferior. En estos países se intuye un fallo estructural profundo, una desconexión entre la preparación del sistema educativo y las demandas reales de sus mercados laborales.

Finalmente, el cuadrante inferior derecho está destinado a aquellos países que, a pesar de una inversión masiva, no logran alcanzar resultados de salida acordes al gasto realizado. Resulta revelador que la gran mayoría de los países analizados han evitado este cuadrante, con la excepción parcial de Alemania (DEU). El sistema alemán presenta un score de entrada muy elevado, comparable al británico, pero su score de salida se sitúa por debajo de la media de la muestra. Este posicionamiento sugiere una posible falta de eficiencia, donde el capital invertido no se traduce de forma directa o proporcional en indicadores de equidad o inserción laboral óptima en comparación con sus vecinos geográficos. Sin embargo, resulta revelador como en el contexto europeo el dicho cuadrante está prácticamente vacío. En un contexto social caracterizado por la implantación definitiva de modelos de Estado democráticos, es lógico que los países, sean mejores o peores sus resultados, no padezcan graves ineficiencias o desconexiones entre su modelo educativo y su mercado laboral, dejando patente un claro sistema de búsqueda de equilibrio, que parece poder dilucidar una tendencia general a la búsqueda de la igualdad de oportunidades.

En términos absolutos, la clasificación del score conjunto de ambos países quedaría de la siguiente forma:

País	Score Entrada	Score Salida	Score Global
LUX	40.23	82.29	**61.26**
FRA	59.31	59.06	**59.18**
GBR	55.08	61.99	**58.53**

CHE	46.19	67.52	**56.86**
LTU	41.57	71.54	**56.55**
MLT	46.67	65.50	**56.08**
NLD	49.45	62.61	**56.03**
BEL	45.94	65.39	**55.67**
DEU	55.27	54.33	**54.80**
FIN	42.96	66.01	**54.49**
SVN	41.35	66.08	**53.72**
SWE	48.63	57.16	**52.90**
DNK	45.36	59.83	**52.59**
ROU	43.11	59.52	**51.32**
NOR	44.96	57.52	**51.24**
HRV	43.81	57.90	**50.85**
LVA	41.83	59.33	**50.58**
IRL	43.35	56.88	**50.12**
POL	43.04	56.98	**50.01**
BGR	44.05	53.93	**48.99**
CZE	41.09	52.66	**46.88**
EST	42.78	50.96	**46.87**
HUN	42.36	51.16	**46.76**
SVK	41.35	49.11	**45.23**
AUT	43.48	46.71	**45.10**
PRT	42.18	46.04	**44.11**
ESP	42.16	45.19	**43.68**
GRC	38.58	31.55	**35.07**
ITA	36.16	31.86	**34.01**

Tabla 20: Ranking de países según el score global

Para comprender qué variables de entrada afectan más a las de salida, procedimos a hacer otro Random Forest de características similares al previamente realizado y que desglosase qué datos tienen más peso a la hora de clasificar un país en función de su desempeño meritocrático. Los resultados son los siguientes:

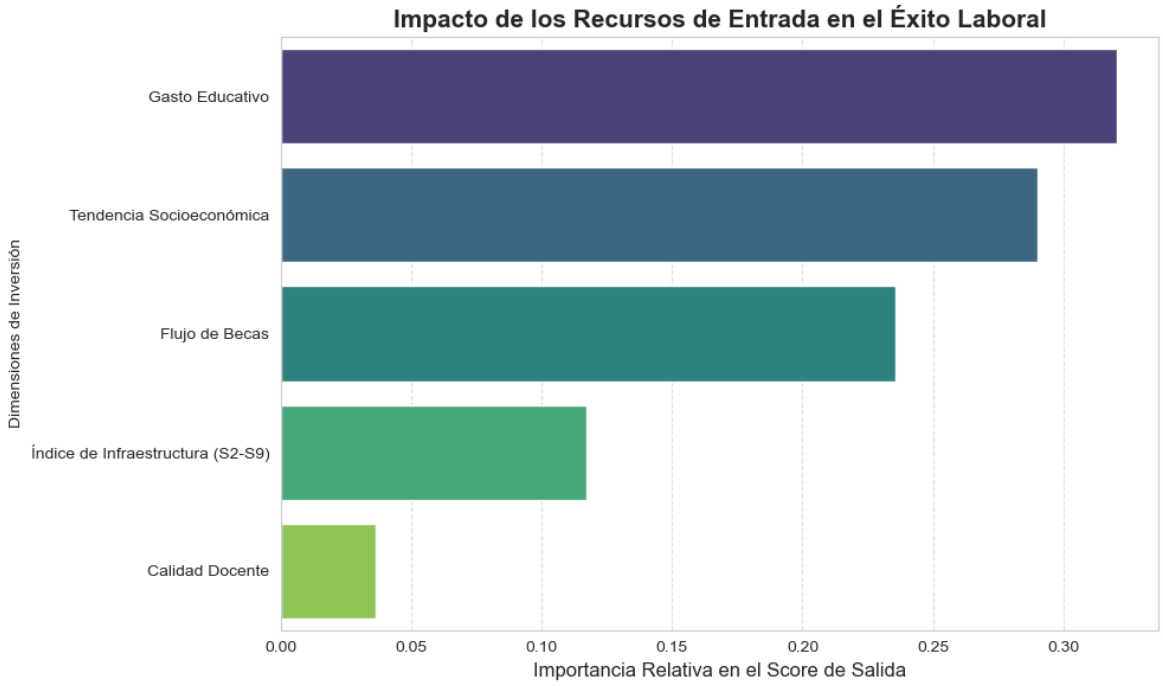


Tabla 21: Peso de las variables de entrada en las variables de salida

En primer lugar, parece predominar el gasto educativo como el factor de mayor relevancia, con una importancia relativa que supera el 30%. Este resultado es de gran importancia, ya que sugiere que, si bien otros recursos específicos parecen ayudar a caracterizar a los países, el volumen total de inversión se intuye como el factor más sólido para explicar niveles elevados de inserción laboral y equidad salarial a nivel agregado.

Acto seguido, aparece el peso del entorno socioeconómico como el segundo motor de éxito con un impacto cercano al 29%. Esta información parece indicar que los resultados laborales están fuertemente condicionados por el contexto de partida de los estudiantes, una conclusión que puede resultar crítica para el diseño de políticas públicas orientadas a la equidad. En este punto, podría surgir la mayor crítica al modelo angular de la investigación, puesto que parece desmentir la idea de una equidad pura en las sociedades.

Por otro lado, el flujo de becas se parece consolidarse como un recurso estratégico de alta eficiencia. Aunque en este modelo ocupa el tercer lugar en importancia, su peso sigue siendo considerablemente elevado. Este dato sugiere que las becas no solo actúan como un rasgo distintivo de los sistemas educativos exitosos, sino que pueden llegar a tener una relación directa en la mejora sustancial de los indicadores de salida.

Finalmente, las variables denominadas de soporte, como el índice de infraestructura y la Calidad Docente, muestran una influencia menor en la predicción directa del score de salida. Esta menor relevancia relativa no parece implicar una carencia de valor, sino que sugiere que su impacto en el mercado laboral es de naturaleza más indirecta o que su contribución ya se encuentra parcialmente capturada por el nivel de gasto educativo general del país.

En otro orden de cosas, podemos observar cómo países como España, Italia y Grecia presentan un patrón caracterizado como "Bajo Retorno", en el que ni el nivel de inversión educativa ni la infraestructura institucional disponible consiguen traducirse de manera eficiente en una inserción laboral sólida o en una reducción significativa de la brecha salarial. Este fenómeno parece apuntar a la existencia de una desconexión profunda entre el sistema educativo y el funcionamiento real de los mercados de trabajo, lo que limita el impacto de las políticas orientadas a la igualdad.

Al situarlo en el contexto geográfico del sur de Europa, esta problemática adquiere una dimensión aún más interesante. Se trata de economías con características estructurales comunes: altos niveles de desempleo juvenil, una fuerte segmentación del mercado laboral, contratos de alta rotación, elevada temporalidad y una presencia significativa de sectores económicos con menor valor añadido. Estas condiciones dificultan que el capital humano generado por el sistema educativo encuentre canales efectivos de integración y desarrollo profesional, alterando la realidad de los datos que manejamos.

Además, parece que los factores institucionales y culturales también juegan un papel relevante. La rigidez de ciertos marcos laborales, la menor conexión entre el sistema educativo y el tejido empresarial, y la persistencia de redes informales en los procesos de contratación pueden reducir la eficacia meritocrática y reforzar desigualdades preexistentes. A esto se suma, en algunos casos, una menor capacidad de absorción de talento cualificado, lo que puede derivar en fenómenos como la sobre cualificación o la emigración de jóvenes formados hacia otros países europeos con mercados laborales más dinámicos.

CONCLUSIONES

Interpretación de resultados

Los resultados del modelaje nos permiten inducir ciertas conclusiones a cerca de la posible relación entre lo que hemos venido considerando educación de calidad y lo que hemos definido como eficacia meritocrática.

En primer lugar, observamos cómo el gasto educativo parece destacar como el factor determinante a la hora de predecir el grado de efectividad con el que la educación en igualdad se traduce en resultados concretos dentro de los mercados laborales. Esto sugiere que no basta únicamente con promover principios de igualdad en el ámbito educativo, sino que resulta importante respaldarlos con una inversión suficiente y sostenida que garantice recursos, programas y políticas adecuadamente implementadas. Un mayor nivel de gasto permite mejorar la calidad de la enseñanza, formar al profesorado en valores inclusivos y desarrollar iniciativas específicas orientadas a reducir brechas de género y otras desigualdades estructurales.

De este modo, los sistemas educativos con mayor financiación parecen seguir una tendencia a generar entornos más equitativos, lo que facilita que dichos valores se reflejen posteriormente en oportunidades laborales más justas, menor discriminación y una participación más equilibrada en el mercado de trabajo.

La tendencia socioeconómica parece constituir el segundo motor de éxito, lo que pone de manifiesto que los resultados en el mercado laboral no parecen depender únicamente de factores individuales o del rendimiento académico, sino que están condicionados por el contexto de origen del alumnado. Este hallazgo evidencia que las oportunidades no se distribuirían *a priori* de manera equitativa desde el inicio, sino que vendrían marcadas por variables como el nivel de ingresos familiar, el capital cultural, el acceso a recursos educativos de calidad y el entorno social en el que se desarrolla el estudiante.

En este sentido, quienes parten de contextos socioeconómicos más favorecidos tenderían a contar con mayores apoyos, redes de contacto más amplias y mejores condiciones para transformar su formación en oportunidades laborales reales. Por el contrario, el alumnado procedente de entornos más vulnerables se enfrentaría a barreras estructurales que dificultarían esa transición, incluso cuando hubieran recibido una educación orientada a

la igualdad. Así, la relevancia de esta variable parece subrayar la necesidad de implementar políticas compensatorias que no solo promuevan la igualdad en el acceso a la educación, sino que también reduzcan las desigualdades de origen, garantizando que la educación pueda actuar verdaderamente como un mecanismo de movilidad social y no como un reproductor de las brechas existentes.

Limitaciones y próximos pasos

La investigación se ha visto condicionada por la periodicidad y los tiempos de publicación de las fuentes estadísticas oficiales, especialmente los informes PISA, cuyos datos no siempre están disponibles con la frecuencia o inmediatez deseadas. Esta limitación ha generado un desfase temporal entre algunas variables de entrada y las variables de salida, dificultando el análisis simultáneo y preciso de las relaciones entre ambas. De este modo, ha sido necesario recurrir a técnicas de interpolación y estimación para completar series históricas incompletas, lo que, si bien permite mantener la continuidad del análisis, introduce un cierto margen de error y reduce el grado de exactitud de los resultados. Además, las distintas bases de datos y su confección entre países y años añade una capa adicional de complejidad, afectando a la comparabilidad internacional y a la robustez de algunas conclusiones.

Por otro lado, el modelo de eficiencia empleado parte del supuesto de que existe una relación relativamente directa entre el sistema educativo y los resultados en el mercado laboral. Sin embargo, esta relación se ve influida por múltiples factores externos que el modelo no puede capturar completamente. Entre ellos destacan las fluctuaciones macroeconómicas, como crisis financieras internacionales o pandemias, que alteran de forma significativa la dinámica del empleo. Asimismo, cambios estructurales en la economía, como la transformación digital o la evolución en la demanda de determinados sectores productivos, pueden modificar las oportunidades laborales disponibles independientemente del nivel educativo de la población. Estos elementos introducen una fuente de incertidumbre que limita la capacidad explicativa del modelo y obligan a interpretar los resultados con cautela, entendiendo que la eficiencia observada no depende exclusivamente de la calidad o equidad del sistema educativo, sino también del contexto económico más amplio en el que este se inserta.

BIBLIOGRAFÍA

Francisco. (2015). *Laudato si'*. Ciudad del Vaticano: Librería Editrice Vaticana. https://www.vatican.va/content/francesco/es/encyclicals/documents/papa-francesco_20150524_enciclica-laudato-si.html

Figueiredo, C. (2025). Conceptualizing 'quality of education': an analysis of European political documents on education. *Frontiers in Education*. https://doi.org/10.3389/feduc.2025.1463412.

Roemer, J., & Trannoy, A. (2016). Equality of Opportunity: Theory and Measurement. *Journal of Economic Literature*, 54, 1288-1332. https://doi.org/10.1257/jel.20151206.

Kim, C., & Choi, Y. (2017). How Meritocracy is Defined Today? Contemporary Aspects of Meritocracy. *Economics & Sociology*, 10, 112-121. https://doi.org/10.14254/2071-789x.2017/10-1/8.

Sandel, M. J. (2021). *La tiranía del mérito: ¿Qué ha sido del bien común?* (A. Santos Montero, Trad.). Debolsillo.

González, A. L., & del Pozo, P. B. (2023). Desigualdad y demografía en áreas metropolitanas de España. *Cuadernos de geografía*, (110), 187-208.

de la Escosura, L. P., & Sánchez-Alonso, B. (2023). El crecimiento económico moderno y su distribución en España. *ICE, Revista de Economía*, (933).

Olarte, C. P. (2024). Desigualdad económica y dominación en España. Treinta años de crecimiento lento y mal repartido (I). *Papeles de relaciones ecosociales y cambio global*, (166), 73-88.

Nosratabadi, S., Atobishi, T., & Hegedűs, S. (2023). Social sustainability of digital transformation: Empirical evidence from EU-27 countries. *Administrative Sciences*, *13*(5), 126.

Bandiera, O., Kotia, A., Lindenlaub, I., Moser, C., & Prat, A. (2024). *Meritocracy across countries* (No. w32375). National Bureau of Economic Research.

Napoletano, T. (2024). Meritocracy, meritocratic education, and equality of opportunity. *Theory and Research in Education*, *22*(1), 3-18.

ANEXOS

https://github.com/pelayo5plaza-lang/Codigo-TFG-Analytics-Pelayo-Plaza.git