PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Facultad de Ciencias Sociales

Data Managment para Finanzas (1FIN57)

Aplicación de métodos de machine learning para la inferencia causal: Análisis de los determinantes del programa de entrenamiento laboral del National Supported Work Demonstration (NSW)

Entrega parcial

Gabriel Sebastián Del Carpio Cuenca

20191565

Lima, 2025

1. Introducción

En las últimas décadas, el análisis causal ha cobrado una creciente relevancia en las ciencias sociales y económicas, particularmente en la evaluación de políticas públicas. Tradicionalmente, este tipo de análisis se ha basado en modelos paramétricos, como la regresión lineal, con supuestos fuertes sobre la elección de variables. Sin embargo, ante el crecimiento exponencial de la información disponible debido al cambio tecnológico de los últimos años, la cantidad de datos, características y dimensiones se ha incrementado, y con ello la complejidad para detectar el tipo de relación entre las variables.

Ante este contexto, el auge de modelos de *machine learning* (ML) ha abierto nuevas oportunidades para abordar problemas de inferencia causal ante contextos de alta dimensionalidad. Aunque estos métodos fueron diseñados con fines predictivos, algunos de sus algoritmos, como los de *random forest* (RF), permiten detectar efectos heterogéneos de tratamiento condicionales a características del conjunto de datos. De tal manera, es posible estimar relaciones causales complejas sin predefinir una forma funcional de las variables.

Esta investigación se propone aplicar y evaluar el desempeño de métodos de ML para la inferencia causal utilizando una base de datos proveniente de un programa de entrenamiento laboral de USA ampliamente estudiado, el National Supported Work Demonstration (NSW). El objetivo principal es comparar la efectividad de diferentes enfoques de machine learning para estimar efectos causales y evaluar si proporcionan mejoras respecto a métodos tradicionales desde el punto de vista de eficiencia y precisión. Con ello, se busca contribuir a la inferencia causal moderna, ofreciendo evidencia empírica sobre el valor añadido de los algoritmos de aprendizaje automático para el análisis de impacto en políticas públicas.

El trabajo se organiza de la siguiente manera: (i), introducción; (ii), justificación del trabajo de investigación; (iii), descripción de los datos; (iv) el análisis de los datos; y (v), primeras conclusiones. Posteriormente se agregarán más secciones y ampliarán los análisis en las mismas. Se finalizará recapitulando los hallazgos, aclarando los riesgos y problemas encontrados en la investigación y estableciendo el plan futuro de trabajo.

2. Justificación

Los métodos de ML están diseñados principalmente para encontrar patrones en los datos con el objetivo de mejorar la predicción. A diferencia de los modelos estadísticos tradicionales, que suelen requerir una especificación funcional explícita, los algoritmos de ML aprenden automáticamente relaciones complejas y no lineales entre variables a partir de grandes cantidades de datos, sin imponer una forma funcional predeterminada.

En el contexto de la inferencia causal, el objetivo no es simplemente predecir un resultado, sino estimar el efecto de una intervención o tratamiento, como el de participar o no en un programa de capacitación laboral. El desafío aquí radica en aislar la relación causal entre el tratamiento y el resultado, controlando por variables de confusión que podrían sesgar la estimación.

Los algoritmos de ML aíslan esa relación a través de cuatro pasos: (i) estimando los parámetros de nuisance (propensidad a ser tratado), (ii) modelando las relaciones complejas a través de validación cruzada, (iii) estimando efectos heterogéneos condicionales a características, y (iv) usando la alta dimensionalidad para encontrar las variables más importantes. Ello permite usar las variables más importantes para estimar su efecto causal en la variable de interés condicional a si las personas fueron apoyadas por la política o no. Esto no solo contribuye a una mejor comprensión del impacto promedio del programa, sino también a identificar subgrupos que se benefician en mayor medida.

Lo anterior es difícil de modelar mediante una regresión lineal, especialmente porque esta requiere satisfacer supuestos como los de identificabilidad, colinealidad perfecta e insesgadez, imposibles de satisfacer en contextos de gran cantidad de datos, con muchos predictores y pocas observaciones.

De esta manera, los métodos de ML contribuyen a la inferencia causal al mejorar la comprensión del impacto promedio de políticas públicas, como los programas sociales, dadas las características más importantes halladas a través de los datos. Además, permiten identificar mejor relaciones complejas entre las variables que los modelos lineales en contexto de alta dimensionalidad.

3. Descripción Data: cualitativa y cuantitativa

Para encontrar tal efecto causal del programa NSW a través de métodos de ML, a continuación, haremos la respectiva descripción de base de datos, así como de las características de las personas que fueron estudiadas. La documentación de esta base de datos está disponible en el siguiente enlace: https://mixtape.scunning.com/05-matching_and_subclassification#example-the-nsw-job-training-program.

Los datos fueron extraídos del experimento del programa de capacitación laboral del NSW realizado a mediados de 70's en USA, en el cual a los sujetos tratados se les garantizó un trabajo durante 9 a 18 meses con salarios modestos sujetos al rendimiento de los sujetos. La base experimental es de 445 observaciones y 10 variables, descritas a continuación:

- treat: sujetos tratados en programa NSW (continua)
- age: edad en años (continua)
- educ: años de educación (continua)
- black: personas de raza negra (booleana)
- hisp: etnicidad hispana (booleana)
- mar: personas casadas (booleana)
- nodegree: personas sin grado profesional (booleana)
- re74: Ganancias reales de 1974 (en miles de dólares)
- re75: Ganancias reales de 1975 (en miles de dólares)
- re78: Ganancias reales de 1978 (en miles de dólares)

Las variables continuas serán abordadas como cuantitativas y las booleanas como cualitativas dadas las finalidades del trabajo.

3.1. Descripción general

A través del script de código SQL adjuntado, observamos que el promedio de edad, años de educación, salarios del 74', 75' y 78' fueron de 25 años, 10 años y 2,103, 1,377 y 5,301 miles de dólares anuales respectivamente. Además, hay 371 personas de raza negra, 39 hispanos, 75 personas casadas y 348 personas sin grado profesional.

Una descripción cruzada nos indica que, en promedio, las personas negras son mayores por un año (25 años) y tienen ingresos del 74' y 75' de 2,115 y 1,298 similares a los 2,037 y 1,775 de los mismos años comparado al grupo que no es blanco. Sin embargo, los ingresos del 78' de personas no blancas de 7,006 superan ampliamente a los 4,961 de las personas negras.

Para el grupo de hispanos, los promedios de edad (23 y 25) e ingresos del 75' son similares a los grupos no hispanos, lo que no ocurre con los salarios del 74' de 1,610 contra los 2,150 los ingresos de los no hispanos y los salarios de 78' de 6,553 contra los 5,180 de los no hispanos; sin embargo, la cantidad de hispanos son mucho menores que la cantidad de 406 personas no hispanas.

Las personas casadas son mayores en cinco años (29) respecto a los no casados en promedio, pero los ingresos de los hispanos son mayores en todos los años, siendo la mayor diferencia en el año 75' y la menor en el 78'.

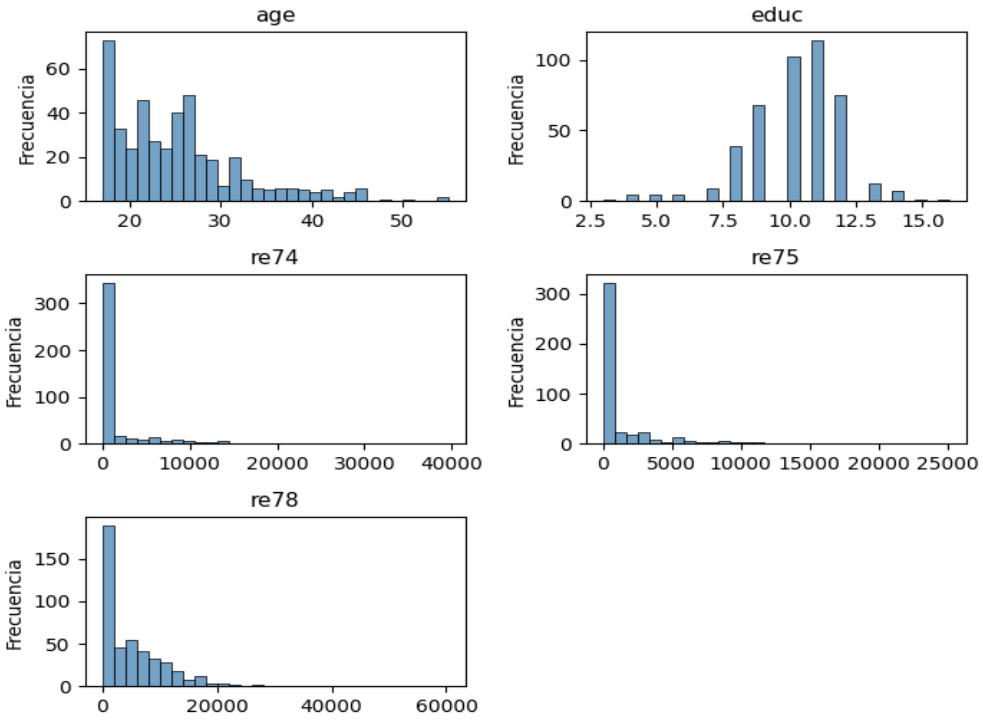
Los años de educación en todos los grupos anteriores son en promedio similares y cercanos a 10, lo que no ocurre lógicamente con las personas con grado

profesional, pues estas tienen en promedio 12 años de educación comparado a los 10 de los que no tienen grado. La diferencia entre los que no tienen grado y los que sí es de 25 a 26 años de edad, 1,910 a 2,791 ingresos del 74', 1,412 a 1,253 ingresos del 75' y 4,930 a 6,632 ingresos del 75'.

En general, la base presenta en su mayoría individuos de edad y años de educación similar, así como una mayor cantidad de personas negras y sin grado profesional mucho mayor a aquellas personas que son hispanas y casadas. La mayor diferencia salarial encontrada por característica fue encontrada en los años 1974 y 1978, lo que puede indicar que probablemente el programa favoreció más a aquellas personas con previa estabilidad económica.

Con el jupyter notebook adjuntado podemos observar ver la estadística y distribución descriptiva de los datos cuantitativos de edad, educación e ingresos anuales. Los ingresos presentan potenciales outliers ya que tienen valores mínimos y máximos muy distantes comparados con la media, lo que refuerza la idea de aplicar métodos de ML para tratar estos problemas.

Distribuciones de variables cuantitativas

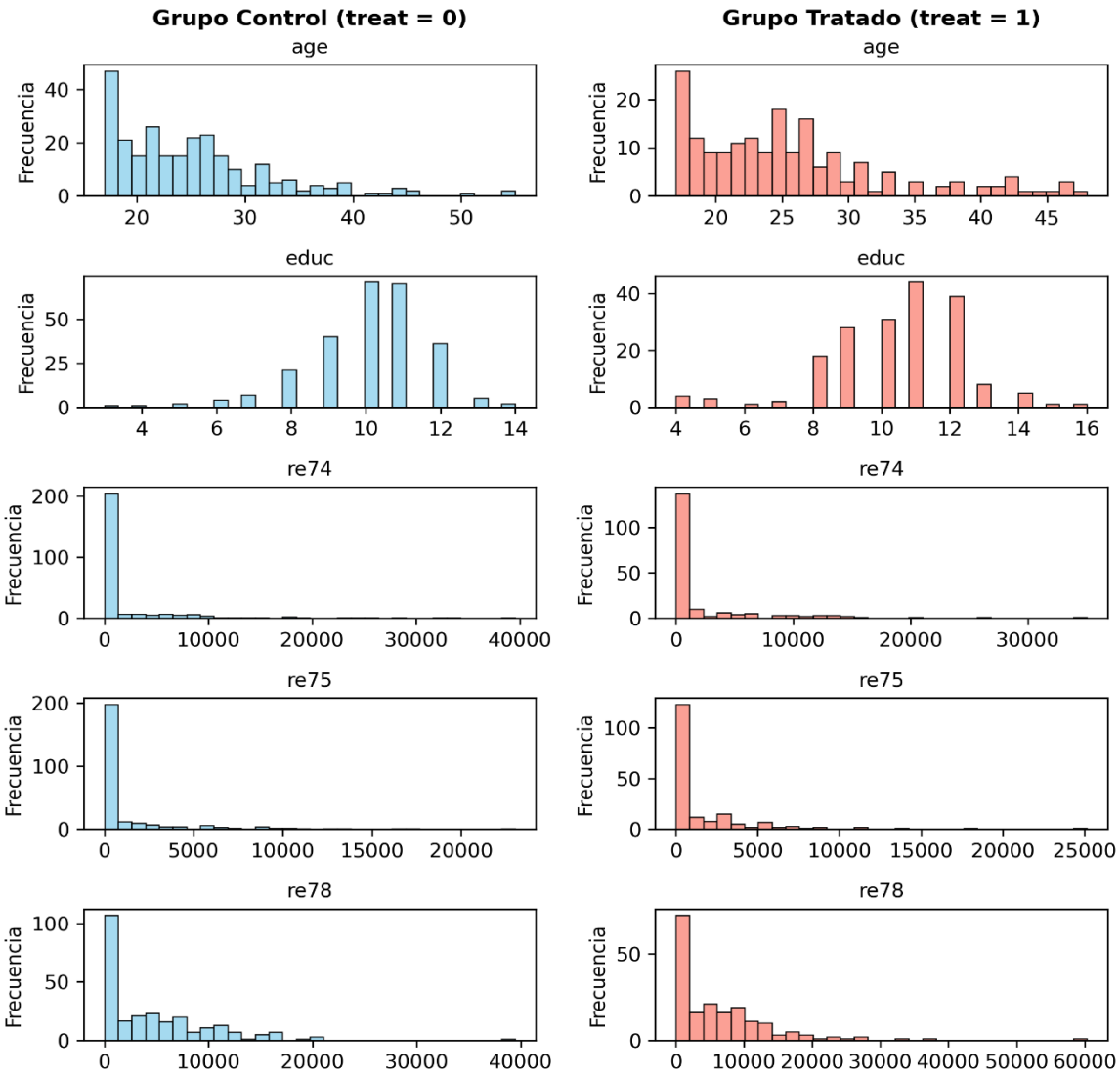


Las distribución de la edad muestra que la mayoría de personas son jóvenes y menores de 30 años. La mayoría de años de educación, como se mencionó antes, esta entre 10 a 12 años. Los ingresos, como es de esperarse, son cercanos a cero, pues son personas sin una fuente de ingresos estable y con poca capacidad para conseguir trabajo.

3.2. Descripción por grupo de tratamiento

Por grupo de tratamiento, las variables cuantitativas muestran promedios similares, siendo las edades del grupo tratado aproximadamente 26, un año más que el promedio del grupo de control. El promedio de años de educación en ambos grupos se aproxima a 10 años. Los salarios promedio si varían, para el grupo de control, fueron de 2 107 en el 74', 1 267 en el 75' y 4 555 en el 78', mientras que de los tratados fueron de 2 096, 1 532 y 6 349, respectiva y aproximadamente.

Distribuciones de variables cualitativas por tratamiento



En ambos grupos, prevalece la mayoría de cantidad personas negras y sin grado, pero la cantidad de hispanos y personas casadas es mayor y menor, respectivamente, en el grupo de control.

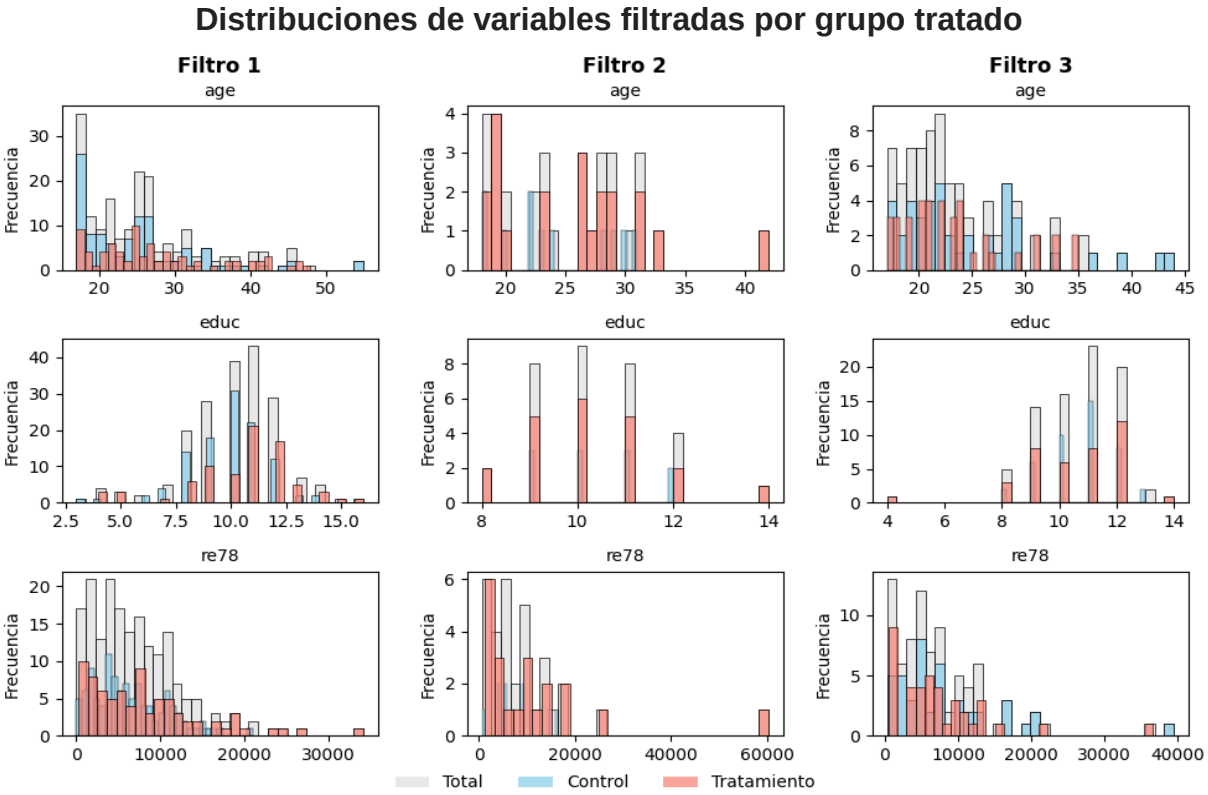
Los histogramas revelan patrones muy parecidos a la base en general. Las mayores diferencias se encuentran en la edad e ingresos de 78'. En el grupo de tratamiento se encuentran personas de mayor edad que el grupo de control. Y los ingresos de 1978 son más distribuidos entorno a los 0 y 20,000 miles de dólares en el grupo de control, mientras que en el de tratamiento se distribuyen más entre 0 y 10,000 miles de dólares.

4. Análisis Data

4.1. Ejercicios de filtrado

Para analizar los datos haremos ejercicios de filtrado entorno a la variable objetivo. Analizaremos si los ingresos anuales aumentaron después de aplicarse el programa de capacitación en 1978. Por la descripción previa, observamos que el mayor diferencial en promedio se presenta en las variables de edad, años de educación e ingresos previos al programa. Por lo anterior, asumiremos los siguientes tres escenarios:

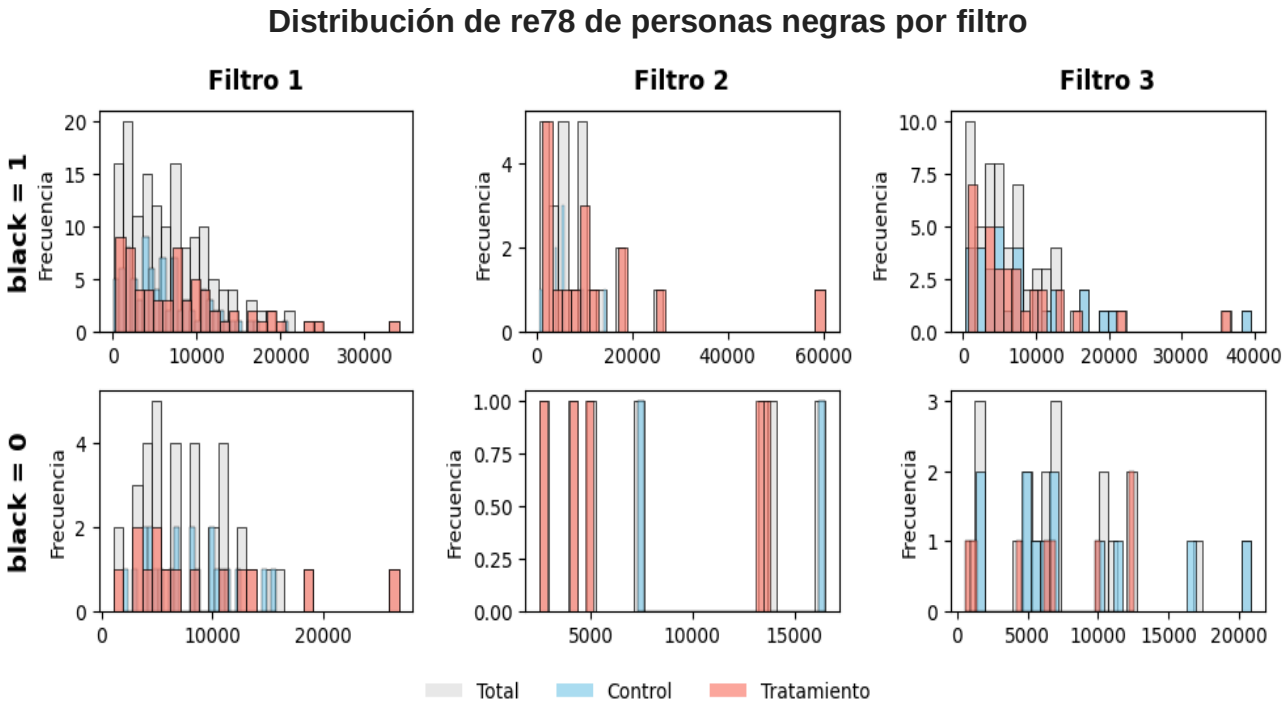
1) Filtro 1: personas que no tuvieron ingresos en 1974 y 1975, pero si en 1978. Este será el grupo desaventajado (GD).
2) Filtro 2: personas que no tuvieron ingresos en 1974, pero si en 1975 y 1978. Este será el grupo sin ventaja (GS)
3) Filtro 3: personas que tuvieron ingresos en los tres años. Este será el grupo aventajado (GA).



Filtrando por el GD observamos que las variables de grupos tratados y de control mantienen la lógica de los datos totales, ello porque, como vimos líneas arriba, la mayor cantidad de personas se ubica dentro de este grupo. Podemos destacar que este GD presenta mayor cantidad de personas jóvenes menores a 30 años y con educación de 10 años comparando la parte de control con la de tratamiento.

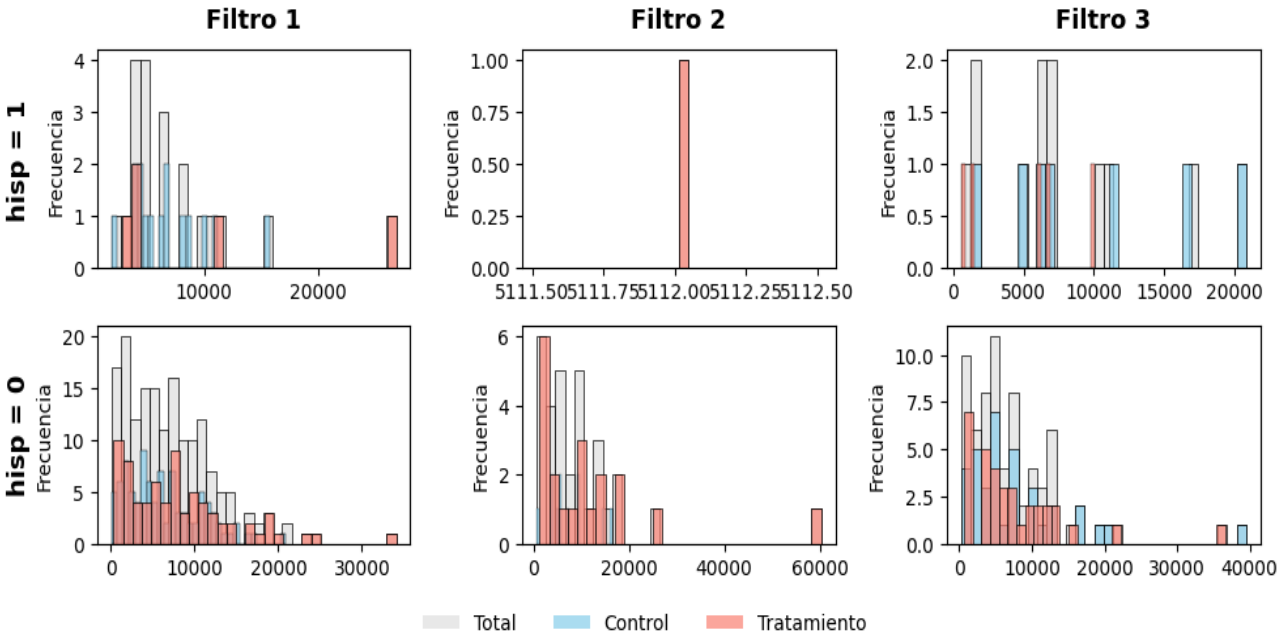
En cambio, el GS presenta la menor cantidad de personas del conjunto de datos. Estas personas tienen edades cercadas a los 20 y 30 años, una educación no menor a los 8 años e ingresos medios más cercanos a los 20,000 dólares anuales después del programa. La mayoría del grupo de tratamiento acapara los datos en el GS.

Los del GA también presenta una cantidad menor de personas que el GD y tienen edades entre los 17 y 25 años, son personas más años de educación (cercano a 12), pero tienen ingresos similares a los del GD, lo que puede ser debido a que este grupo no son el sujeto foto del programa. En este grupo también, los grupos de tratamiento suelen ser más jóvenes y ganar menos que los de control.

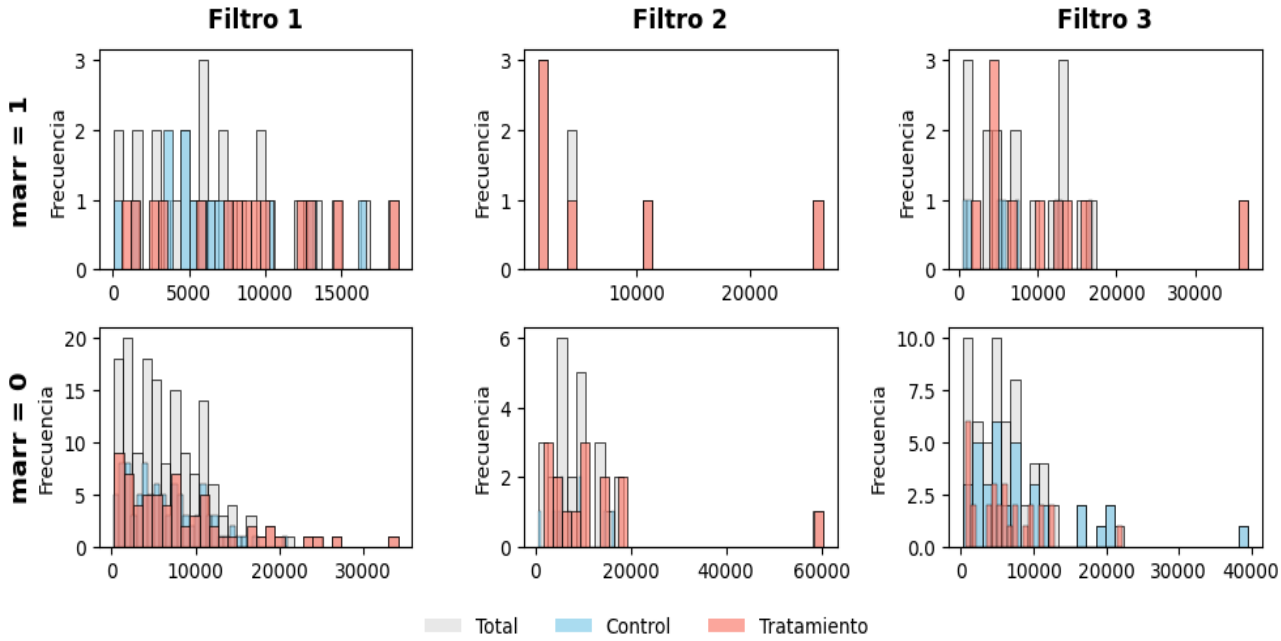


Realizando el filtrado por variable cualitativa y enfocado en los ingresos de 1978, encontramos que los GD, GS y GA tienen en su mayoría a personas negras en su conjunto de datos. En el GD, los ingresos mantienen la lógica inicial en sujetos controlados y tratados. Mientras que hay personas negras con mayores ingresos que el grupo de personas no negras dentro del GS. Esto se repite en el GA, pero el grupo de control tienen más cantidad de personas con ingresos mayores que la cantidad tratada.

Distribución de re78 de personas hispanas por filtro



Distribución de re78 de personas casadas por filtro

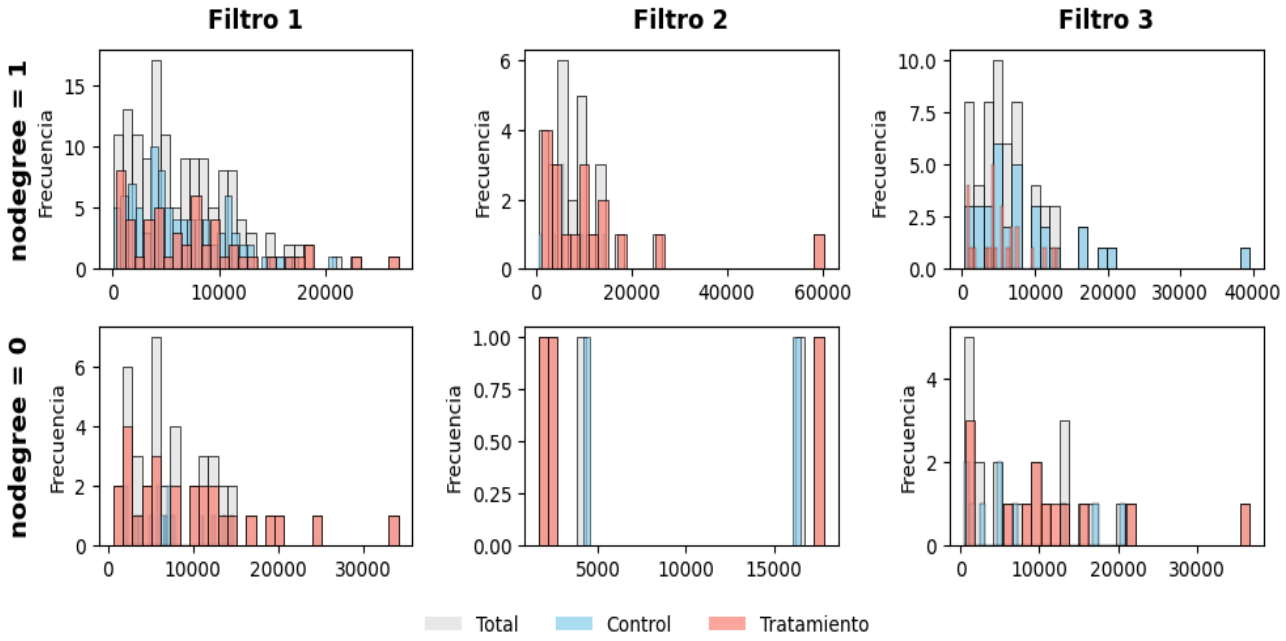


Dado que las personas hispanas y casadas son la menor cantidad de sujetos en el conjunto de datos, la lógica es similar dentro del grupo no hispanos y no casados.

Dentro de las personas hispanas, el GD tiene pocas personas, las cuales tienen ingresos bajos, pero no muchas sin ningún ingresos. El GS solo presenta una persona con un ingreso alto 5,112, la cual pertenece al grupo tratado. El GA tiene personas con ingresos menores a 20,000.

Las personas casadas, en cambio, tiene personas con salarios concentrados entre 5,000 y 10,000 del GD, pocas personas en el GS y una lógica similar en el GA a la lógica del conjunto de datos en general.

Distribución de re78 de personas sin grado profesional por filtro



Analizando a las personas sin grado profesional, dado que también tienen mayoría en el conjunto de datos, la lógica general se mantiene en cada filtro. Lo más resaltante es que el GA tienen más personas en el grupo de control que el de tratamiento.

En resumen, hay una menor cantidad de personas dentro del GS y GA, esto debido a que el programa busca mejorar las condiciones económicas de los grupos más desaventajados, los cuales son aquellos que no tuvieron ingresos previos al programa.

Para complementar el ejercicio de filtrado, describiremos los gráficos de cajas sobre la variable de ingresos de 1978 por variable cualitativa y filtro, ello para identificar lo valores máximos, mínimo y medios, así como los outliers.

Gráfico de cajas de re78 de personas negras por filtro

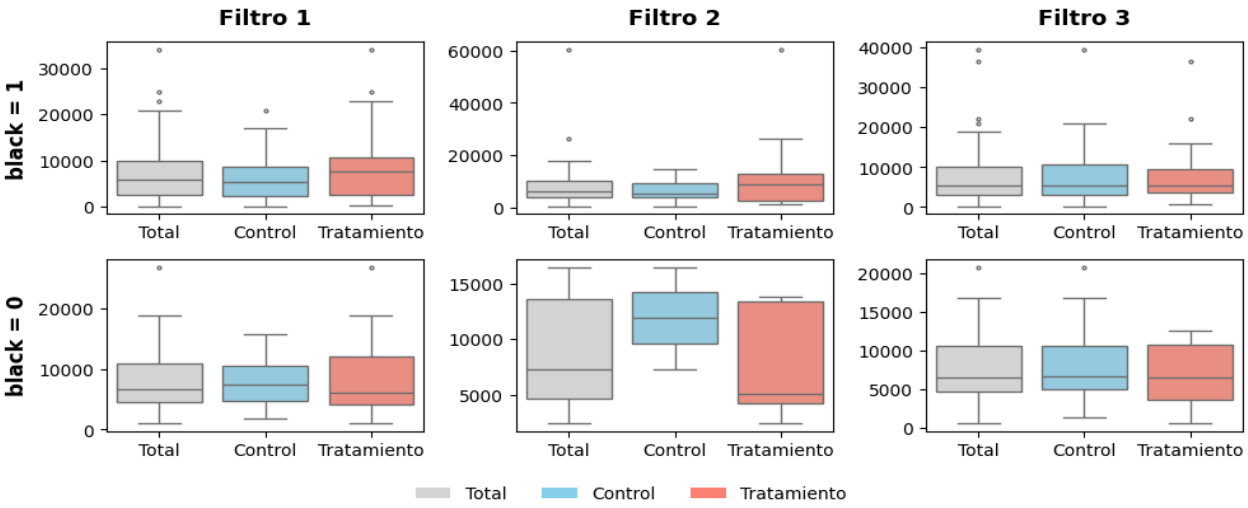


Gráfico de cajas de re78 de personas hispanas por filtro

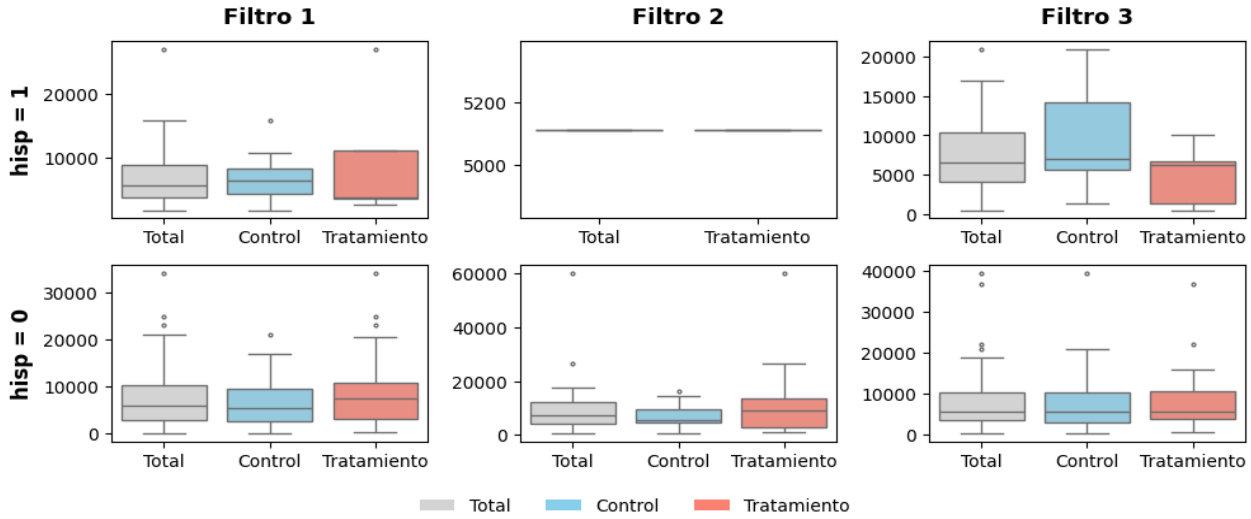
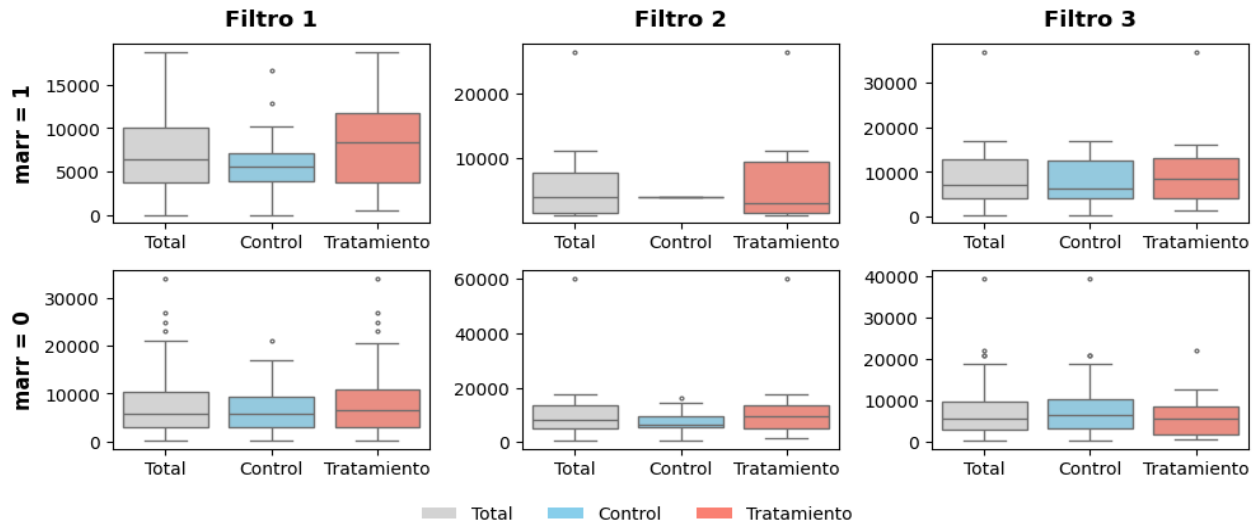
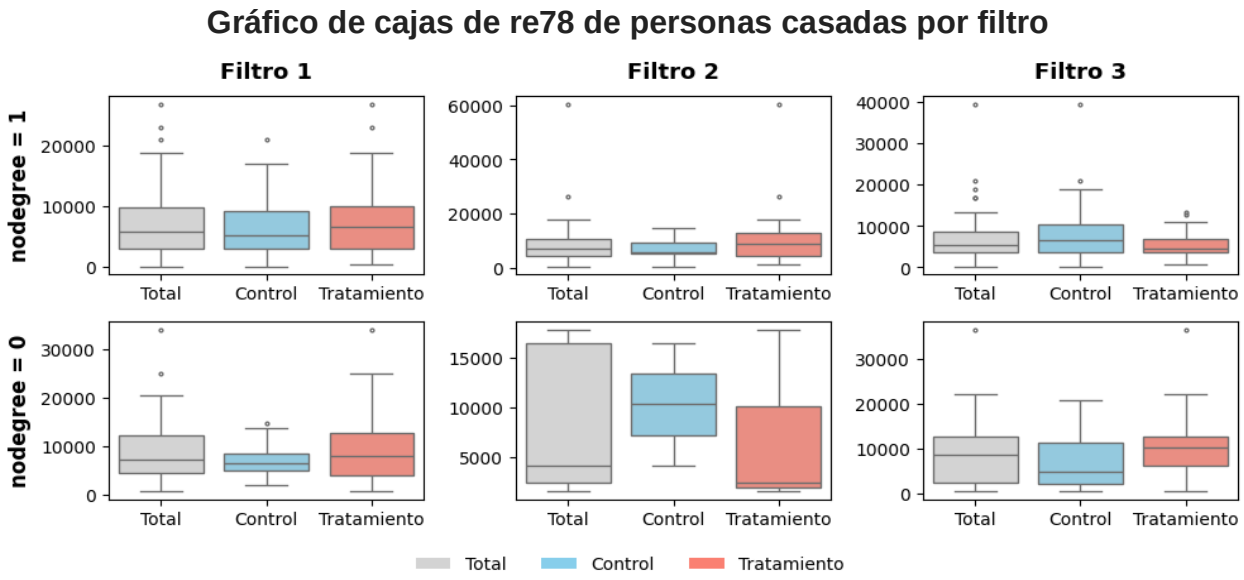


Gráfico de cajas de re78 de personas casadas por filtro





Lo más importante a destacar en el análisis de cajas es que las variables filtradas presentan datos muy por fuera del percentil superior y que el los GD y GA tienen datos distribuidos de manera más concentrada. Solo las personas no negras y con grado profesional tienen datos sin outliers y con ingresos entre 5,000 y 15,000 miles de dólares.

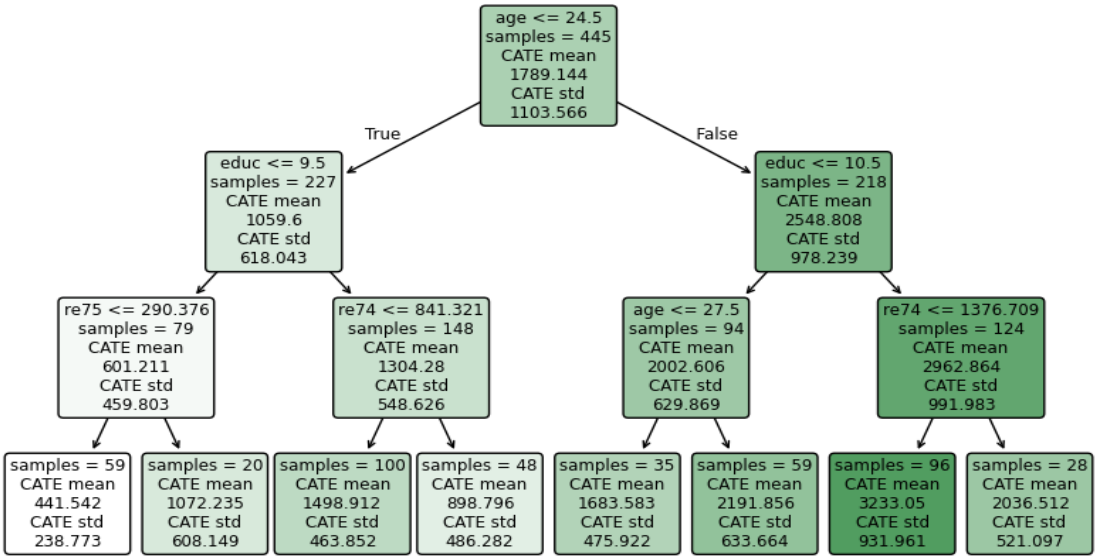
4.2. Aplicación de Causal Forest

Para complementar el análisis de datos, empleamos la variante del modelo RF a uno orientado a la inferencia causal llamado *causal forest* (CF). Este está diseñado específicamente para estimar efectos causales heterogéneos. En lugar de predecir un resultado, busca estimar cómo varía el efecto de un tratamiento según las características de cada individuo.

El CF crea árboles que particionan los datos en subgrupos homogéneos en cuanto al efecto del tratamiento (no en cuanto al resultado). Para evitar sobreajuste, usa como “honestidad” la división los datos en dos subconjuntos: uno donde crea divisiones o “hojas” del árbol y otro para estimar el efecto de tratamiento dentro de cada hoja. Por último, al igual que el RF, repite este proceso múltiples veces y promedia las estimaciones para hallar el efectos promedio de tratamiento condicional (CATE) del bosque.

Para este caso, dado que no tenemos un conjunto de observaciones elevado en relación a las variables, establecemos una cantidad máxima de división de cuatro hojas, es decir, si el algoritmo cree que pueden hacerse más divisiones, serán como máximo cuatro.

Causal Forest y CATE por hoja

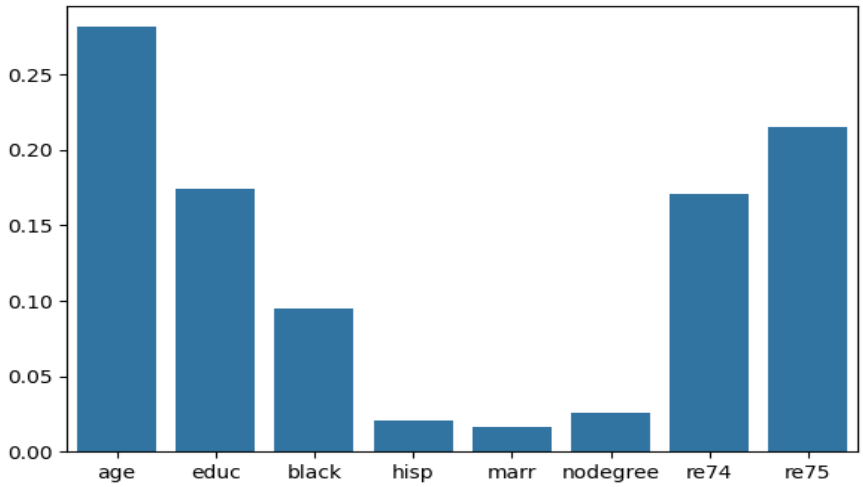


Como resultado del CF, observamos que aquellas personas mayores a 25 años, con más de 11 años de educación y con ingresos menores a 1,377 miles de dólares en 1974 son las que presentan el CATE más elevado de todas las hojas, con un promedio de 3,233 miles de dólares en 1978 luego de aplicarse el programa.

Otro resultado importante, arroja que las personas menores a 25 años, con más de 10 años de educación y con ingresos menores a 841 miles de dólares en 1974 presentan un CATE de 1499 miles de dólares después de aplicarse el programa.

Lo anterior es lo más relevante a destacar dado que ambas hojas presentan la mayor cantidad de la muestra, lo que indica que factores como la edad, educación y el éxito económico del 74' previo al programa son los más determinantes para que los individuos incrementen sus ingresos después de capacitarse laboralmente.

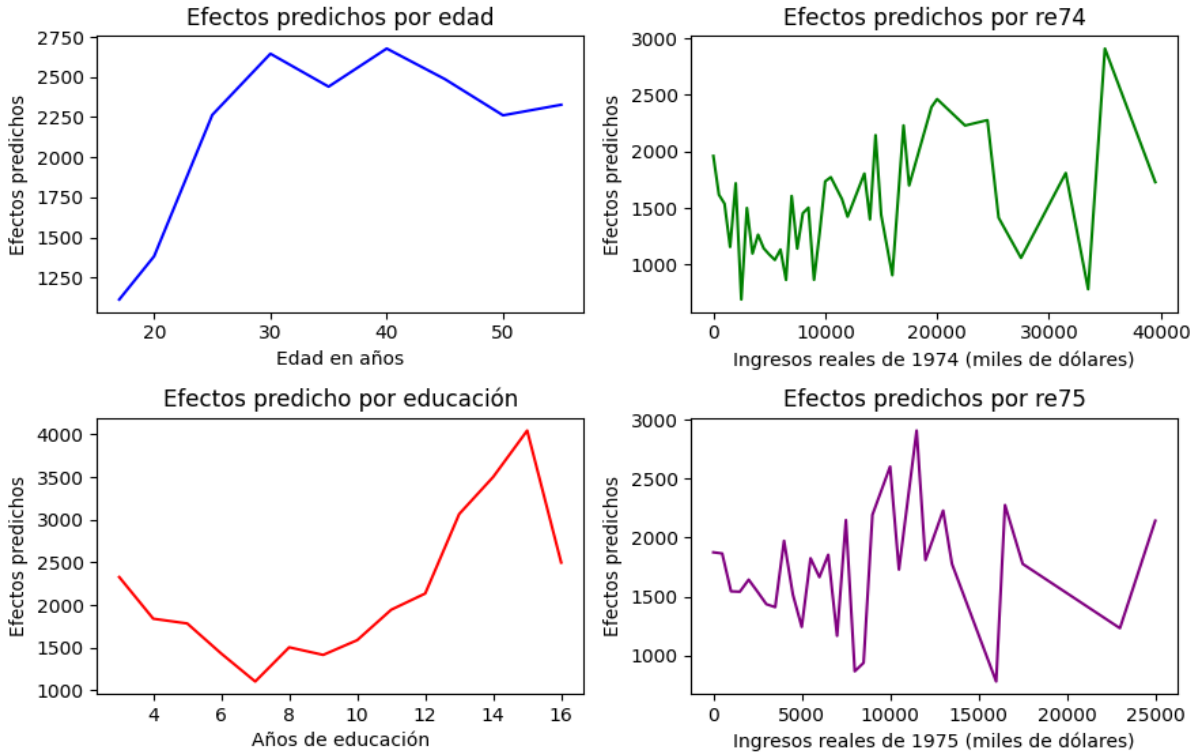
Importancia de los Features



Graficando la importancia de las características estimada por el algoritmo, los resultados coindicen con lo observado, pues las variables con mayor importancia son la edad (0.281), los ingresos del 75’ (0.215), los años de educación (0.174) y los ingresos del 74’ (0.171).

Finalmente, para observar cómo afectan estas variables seleccionadas a la predictibilidad de los ingresos de 1975, graficaremos los efectos heterogéneos predichos por cada variable.

Predicción de re75 por variable importante



Los años de edad y educación muestran efectos no lineales predichos en relación a la variable de interés mucho más claros que las variables de ingresos del 74’ y 75’.

Las personas entre 30 y 40 años son las que más se benefician del programa, mientras más joven o viejo eres es más probable no tener mayores ingresos después del programa. Las personas altamente educadas, hasta antes de los 15 años, se benefician enormemente del programa, en cambio, mientras menos años de educación tengan las personas es menos probable tener mejores ingresos en el 78’. Por último, parece que las personas con ingresos cercanos 35,000 en el 74’ aprovechan más el programa que el resto en el mismo año. El máximo salario predicho por los ingresos del 75’ se da cuando las personas tienen poco más de 10,000 el mismo año.

5. Primeras conclusiones

El análisis descriptivo, el ejercicio de filtrado y la aplicación del CF muestran que la edad, los años de educación y la estabilidad económica previa al programa son las variables más determinantes para que los individuos tengan mayor éxito económico promovido a través del programa NSW.

El análisis indica que las personas que tienden a ser muy jóvenes o muy viejas no suelen ser tan beneficiadas por el programa a comparación de las que no están en ese grupo. Además, los individuos con menos de 12 años de educación (la mayoría del conjunto de datos) son propensos a no beneficiarse del programa. Lo anterior puede ocurrir porque, al ser personas jóvenes, y probablemente con pocos años de educación, no tienen la suficiente experiencia para encontrar mejores trabajos. Por lo contrario, las personas más viejas, si bien se benefician por la experiencia, pueden generar menores ingresos ya que estos dependen del rendimiento, el cual suele disminuirse con la edad.

Los ingresos anuales de 1974 y 1975 suelen tener efectos diversos porque típicamente los ingresos dependen de otras variables como la raza, la etnicidad, el matrimonio y título profesional. Sin embargo, parece que las personas con éxito económico elevado al promedio en 1974 aprovechan mejor el programa que los que no tienen tal “ventaja”.

Si bien los ejercicios filtrados tenían el objetivo de establecer relaciones preliminares bajo circunstancias relacionadas a ciertas ventajas y desventajas económicas condicionadas a las variables cualitativas, se encontraron otros hallazgos. Principalmente, el GA es el que, si bien presenta pocos individuos, son los que tienen mayores ingresos anuales en 1978 en comparación a los GD. Una conclusión similar se mantiene cuando se aplica el filtro de GS a las variables cualitativas.

Un factor delimitante a aclarar es que al separar los individuos en grupos de tratamiento y control de manera aleatoria ciertas estimaciones pueden estar sesgadas ya que no existe una cantidad de personas homogénea en ambos grupos. Por lo anterior, posteriormente se ampliará el análisis de datos a conjuntos observacionales que provienen de la misma fuente. La lógica detrás de esto es que el grupo de tratado se mantenga igual pero el de control sea externo, haciendo los resultados más robustos y consistentes.