FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

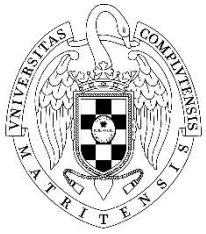
Curso 2018/2019

Trabajo de Fin de Máster

TITULO: Estimación indirecta de la Discriminación Salarial de Género en España con técnicas de Minería de Datos. Metodología Oaxaca-Blinder.

***Alumno*: Raquel García Ballesteros**
***Tutor*: Lorenzo Escot Mangas**

Noviembre de 2019



UNIVERSIDAD COMPLUTENSE
MADRID

Contenido

Índice de tablas e ilustraciones.

Tablas

Ilustraciones

I. Introducción.

1. Motivación inicial y punto de partida de la investigación: identificación del problema de estudio y estado del arte.

En este Trabajo de Fin de Máster se pretende abordar una estimación indirecta de la discriminación salarial de género en España con diferentes modelos a partir de diferentes técnicas de minería de datos usando la metodología de Oaxaca-Blinder. La discriminación salarial de género es parte de la brecha salarial de género.

Las diferencias salariales por hora entre mujeres y hombres no se pueden interpretar directamente como discriminación de género, porque ambos colectivos presentan una distribución diferente de las variables que el mercado retribuye de forma diversa. Hecho que se va a comprobar más adelante cuando se obtiene unas ecuaciones de salarios diferentes para los hombres y para las mujeres.

La principal variable explicativa de la brecha salarial es la segregación laboral en sus diferentes dimensiones: la distribución de las responsabilidades familiares, los diferentes usos del tiempo, el peso del trabajo a tiempo parcial, la desigual participación en el mercado laboral, la concentración en sectores y empresas con peores tipos de contrato y salarios. La segunda variable explicativa es el impacto de la asimetría al acceso de complementos salariales por parte de las mujeres, que ayudan a comprender no sólo la naturaleza de las diferencias salariales de género en España, sino también el mecanismo por el que se producen.

La discriminación salarial de género no es sólo un fallo del mercado laboral, sino que supone una pérdida de capacidad productiva (mano de obra femenina) y de bienestar social hacia la mujer y la sociedad en su conjunto.

La desigualdad salarial tiene importantes implicaciones sobre la eficiencia económica. Según Becker (1975) y Arrow (1973) dos trabajadores de distinto sexo, pero con la misma dotación de capital humano son sustitutos perfectos en la producción, por lo tanto, cualquier diferencia salarial es discriminatoria.

Este trabajo intenta abordar el problema de la discriminación salarial de género en España con los últimos datos disponibles hasta la fecha. El último dato se corresponde con el año 2018, publicado el pasado 27 de junio de 2019 para la Encuesta de Condiciones de Vida (ECV).

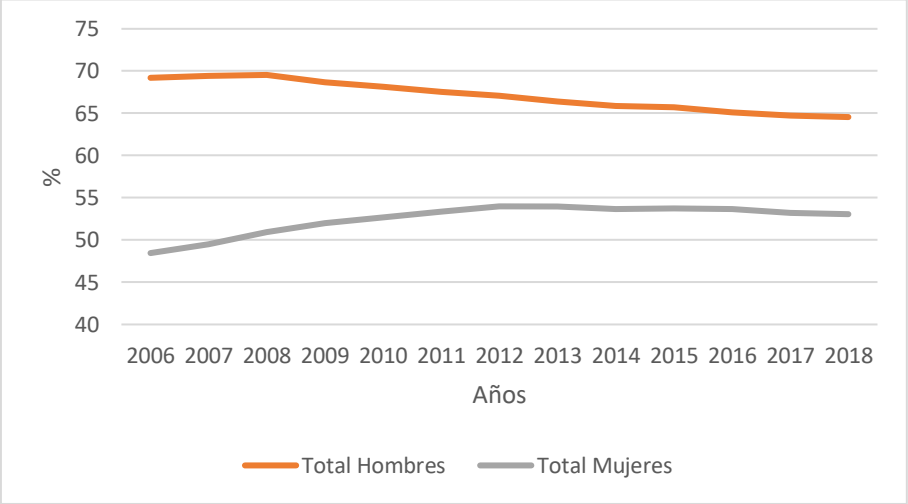
Uno de los principales objetivos que busca este trabajo es analizar algunas de las causas últimas de la desigualdad de género en el mercado de trabajo.

Principalmente el trabajo se centra en la estimación de un índice de discriminación salarial de género por la metodología de Oaxaca – Blinder aplicada a las predicciones de salarios obtenidas por regresión lineal, redes neuronales, árboles de regresión, bagging, random forest y gradient boosting para poder ver cual modelo se ajusta mejor a la realidad.

Antes de ello tenemos que ponernos previamente en situación en el panorama del mercado laboral español. Para ello contamos con la ayuda de los siguientes gráficos[1] que ayudan a ver la participación de la mujer en el mercado laboral. La participación de la mujer en el mercado laboral se puede estimar con la tasa de actividad, la tasa de empleo (ocupación) y la tasa de paro.

1.1 Las tasas de actividad, empleo y paro femenina y masculina.

Ilustración 1. Tasa de Actividad en España (2006-2018)



Fuente: Elaboración propia a partir de datos de la EPA (INE).

En la Ilustración 1 está representada la tasa de actividad. A grandes rasgos se puede observar el interés creciente de la mujer por introducirse en el mercado laboral en el período 2006-2012 y que se mantiene más o menos estable en el resto de los períodos. Por el lado de los hombres se observa una tendencia decreciente de la tasa de actividad. El gráfico muestra un embudo de convergencia lenta entre ambas tasas.

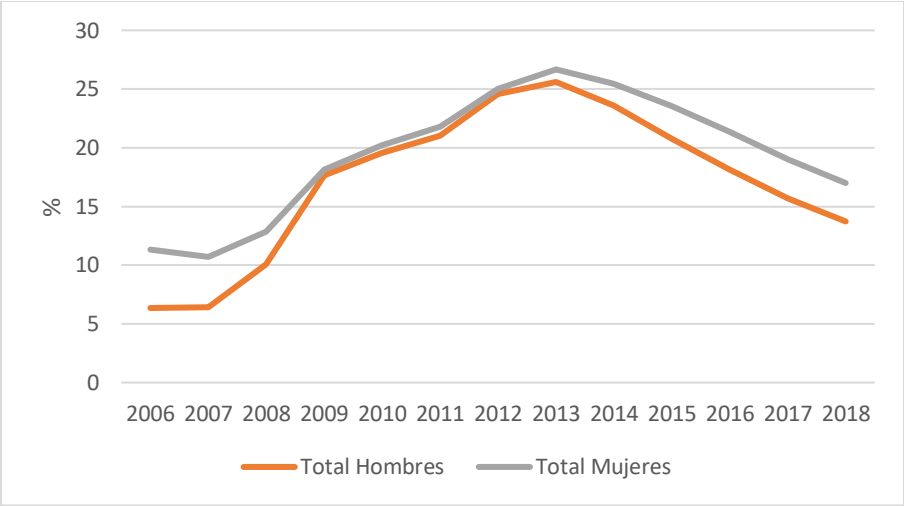
Podemos hablar, por tanto, de una reducción de las diferencias existentes de género en la búsqueda activa de un puesto de trabajo.

Hay que destacar que la tasa de actividad, la decisión de incorporarse al mercado de trabajo, no se ve tan afectada por el contexto macroeconómico como por la circunstancias microeconómicas y sociales del entorno del individuo. Las personas son libres de elegir trabajar o no.

Esta última idea explica porque no ocurre lo mismo con la tasa de desempleo y de ocupación. Para estas existe una interacción necesaria entre la oferta y la demanda de empleo. Esta relación a su vez se ve influenciada por el ciclo económico.

[1] Los datos con los que se han elaborado los gráficos provienen de la Encuesta de Población Activa (EPA) del Instituto Nacional de Estadística (INE). Se han trabajado con los datos anuales que son calculados como la media de los cuatro trimestres del año. La tabla que contiene las cifras de cada una de las tres tasas para cada período y género se pueden consultar en el Anexo 1.

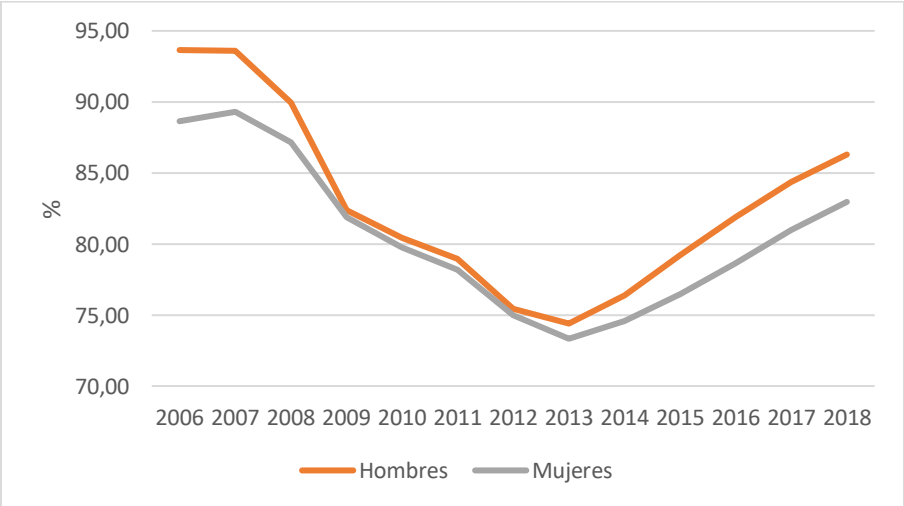
Ilustración 2. Tasa de Paro en España (2006-2018)



Fuente: Elaboración propia a partir de datos de la EPA (INE).

En la Ilustración 2 se encuentra representada la Tasa de Paro para ambos géneros. La tasa de paro para ambos géneros tiene forma de parábola invertida. Durante el período de tiempo de 2006 a 2013 ambas tasas son equiparables al alza, aunque siempre la tasa de desempleo femenina es superior a la masculina. La tendencia creciente en el intervalo de tiempo mencionado se debe en parte a la composición sectorial de los trabajadores en el mercado de trabajo español y al desigual reparto de los efectos de la crisis por actividad económica. Además del período de crisis económica sucedido durante 2007-2013. A partir de 2013 se produce un descenso importante en ambas tasas, siendo la femenina superior a la masculina, es más, la masculina ha descendido mucho más en los últimos años que la femenina.

Ilustración 3. Tasa de Ocupación en España (2006-2018)



Fuente: Elaboración propia a partir de datos de la EPA (INE).

Por último, en la Ilustración 3 está representado la Tasa de Ocupación, o también conocida como Tasa de Empleo. Esta tasa al contrario de la de desempleo tiene forma de parábola (la de desempleo era invertida) y nos muestra que los hombres tienen una tasa de empleo mayor que la de las mujeres. Pero los hombres tienen más

irregularidades en su evolución en el tiempo, pero con una forma muy parecida a la de las mujeres.

1.2 Estado del arte.

Existen tres teorías fundamentales para explicar las diferencias salariales y ocupacionales entre los hombres y las mujeres en el ámbito laboral que parten de supuestos distintos sobre el funcionamiento del mercado de trabajo.

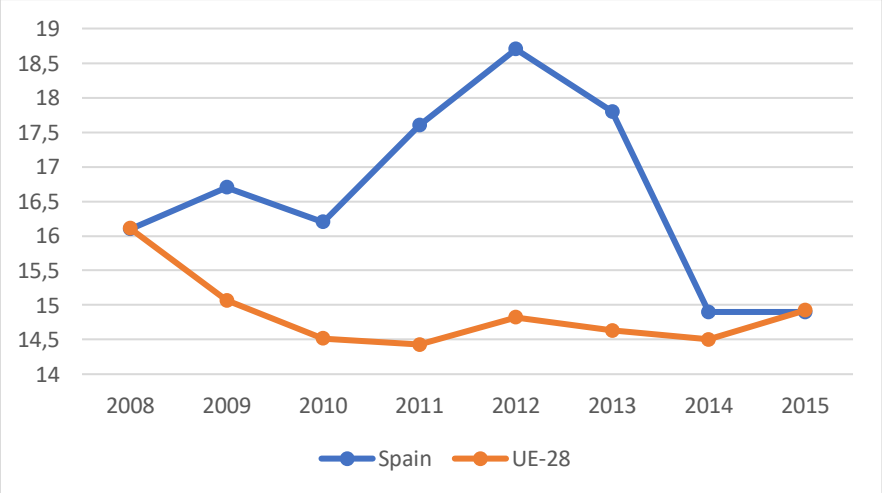
1. Teoría neoclásica del capital humano. Establece una relación causal entre la educación, la productividad y las ganancias. De este modo las diferencias salariales vendrán explicadas por diferencias en productividad.
2. Teoría Institucionalista (Doeringer y Piore, 1971). Sostiene que el mercado de trabajo no funciona competitivamente por la existencia de restricciones institucionales que explican la conducta de los distintos agentes económicos. Se observa directamente el funcionamiento del mercado.
3. Teoría Marxista (Bravermann, 1974). Supone que la relación laboral es una relación de mercado y una relación entre dos tipos de personas, o clases, con intereses distintos y a menudo contrapuestos en el proceso productivo. El análisis de clases podría explicar la estructura de salarios y las características relevantes para el mercado de trabajo según los marxistas.

Cada una de estas hipótesis ayuda a explicar una parte de las diferencias salariales y ocupacionales por razón de sexo en el mercado laboral. Pero nos vamos a centrar en la Teoría del Capital Humano, que se viene aplicando en todos los estudios sobre la misma temática que aborda este trabajo para el caso de España.

2. Justificación del proyecto.

El panorama de la brecha salarial de género en España se puede ver resumido de forma visual en estos datos. Se puede ver como en España en el período 2008-2015 la tendencia ha sido la disminución de la brecha salarial, pero no su supresión. Se encuentra en línea con la Unión Europea de los 28.

Ilustración 4. Evolución de la Brecha Salarial de género por hora 2008-2015. España comparada con la UE-28



Fuente: INE y Eurostat. Brecha salarial de género.

En la Tabla 1 se encuentra de manera resumida las diferentes brechas que se suceden hasta la obtención del salario neto.

Un hombre y una mujer de la misma categoría profesional tienen el mismo salario base, establecido por ley, si no se trataría de discriminación directa. Se determina de forma automática según la categoría ocupacional del trabajador.

Sin embargo, los trabajos femeninos se penalizan fundamentalmente a través de los complementos salariales por su carácter discrecional.

Al desagregar el salario total en salario base y complementos salariales obtenemos dos resultados muy importantes:

En España no se cumple "mismo pago por mismo trabajo" pues, al comparar el colectivo de mujeres con el de hombres, nos encontramos con una clara diferencia salarial del 22.53 % a favor de los hombres en salario bruto, según los datos de la Encuesta de Estructura Salarial del 2014 proporcionados en la Tabla 1.

La diferencia en complementos salariales es del 30,44% a favor de los hombres, pero también los demás componentes del salario, distintos del salario base (nocturnidad, horas extra, pagas extra), son susceptiblemente más altos en los hombres que en las mujeres. Este hecho se debe en parte al mayor poder de negociación colectiva en los sectores masculinizados.

La diferencia salarial en horas extraordinarias es por la diferencia de jornada laboral, puesto que el colectivo femenino está empleado en mayor proporción en jornadas a tiempo parcial y en horarios de mayor compatibilidad familiar.

Tabla 1. Brecha salarial de género (en puntos %) de acuerdo con la estructura del salario bruto mensual por sexo. España 2014.

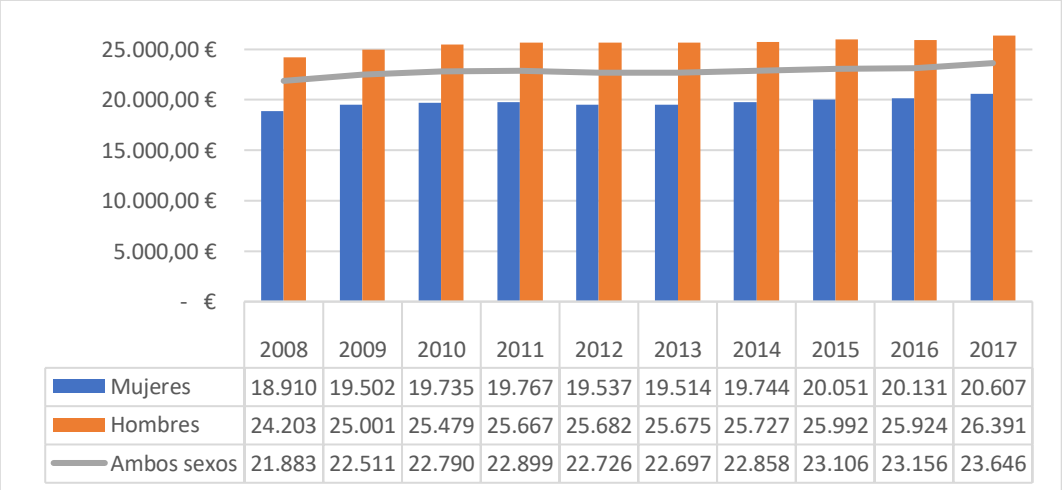
	BRECHA
Salario base	17,99
Complementos salariales	30,44
Complementos salariales por razón de nocturnidad, turnicidad o trabajo durante el fin de semana	33,03
Pagos por horas extraordinarias	78,88
Salario ordinario	22,28
Pagas extraordinarias	36,07
Salario bruto	22,53
Contribuciones a la Seguridad Social a cargo del trabajador	19,91
Retenciones IRPF	34,60
Salario neto	20,16

Fuente: INE, Encuesta de Estructura Salarial de 2014.

En los siguientes gráficos se muestra la evolución reciente de los salarios brutos anuales de varones y mujeres en España, a partir de los datos de la Encuesta de Estructura Salarial (EES). De su análisis se pueden destacar los siguientes aspectos:

1) El salario medio bruto anual (nominal) para ambos sexos sigue una tendencia creciente a lo largo de todo el período 2008-2017. Pasando en España de 21.883 euros en 2008 a 23.646 euros en 2017. Si se observa de forma individual a cada uno de los sexos también se ve que la tendencia creciente se extrapola a cada uno de los sexos: las mujeres han pasado de 18.910 euros en 2008 a 20.607 euros en 2017 y, los hombres han pasado de 24.203 euros en 2008 a 26.391 euros en 2017.
2) El salario medio bruto anual (nominal) de los varones ha permanecido por encima del de las mujeres durante todo el período. De esta forma, la brecha salarial entre varones y mujeres ha sido alta.
3) Analizando la ratio mujer/hombre (cociente entre el salario medio anual de las mujeres entre el de los hombres), también se observa que comparando 2008 con 2017 ha disminuido en 0.05 puntos porcentuales. Se ha pasado del 78.13% al 78.08%. Lo que más llama la atención es la disminución que se produjo en el período 2008 a 2013 fruto de la crisis económica. La llegada de la crisis económica pudo tener un doble efecto sobre la brecha salarial (Murillo y Simón, 2013): la destrucción de empleo asociada a la crisis ha afectado en mayor medida al colectivo de varones y también al colectivo de trabajadoras y trabajadores menos cualificados, y esto podría haber supuesto tanto una reducción de la brecha salarial media como un aumento, dependiendo del efecto final del colectivo relativamente más afectado por los ajustes salariales y la expulsión del mercado haya sido el de las mujeres o el de los hombres. Pero después, se volvieron a recuperar los salarios entre ambos sexos, lo que explica el incremento de la ratio, tomando la forma de parábola. Parece que en los últimos años ha habido una tendencia general hacia el incremento de la ratio mujer/varón en el salario anual.
4) La brecha salarial de género presenta una tendencia claramente decreciente después de la crisis económica.

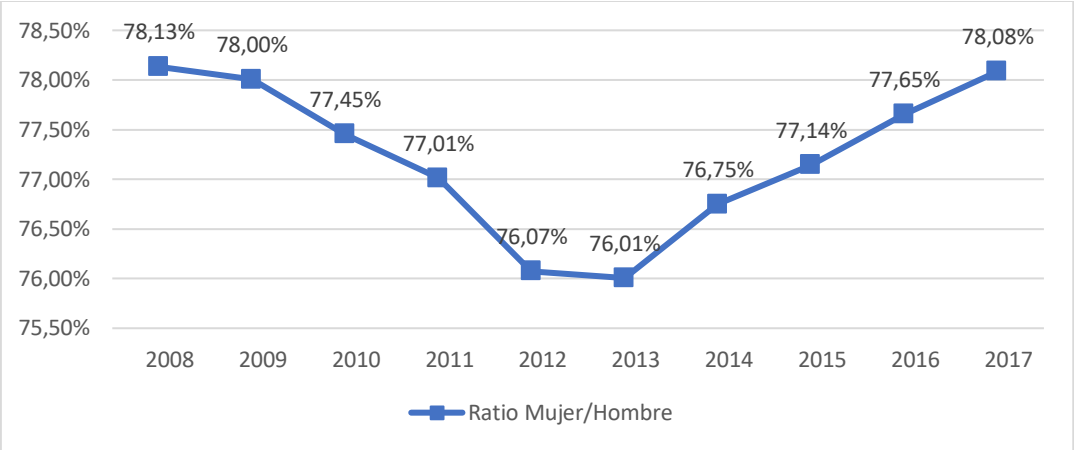
Ilustración 5. Evolución de los Salarios Brutos Anuales por Sexo (2008-2017)



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Mujeres	18.910	19.502	19.735	19.767	19.537	19.514	19.744	20.051	20.131	20.607
Hombres	24.203	25.001	25.479	25.667	25.682	25.675	25.727	25.992	25.924	26.391
Ambos sexos	21.883	22.511	22.790	22.899	22.726	22.697	22.858	23.106	23.156	23.646

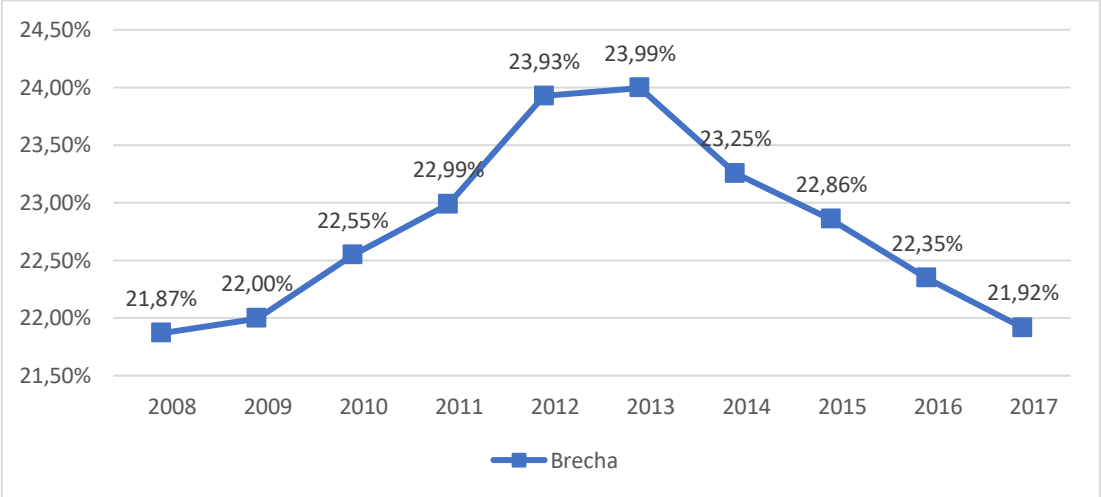
Fuente: Encuesta de Estructura Salarial (EES), INE.

Ilustración 6. Ratio mujer/hombre en España (2008-2017)



Fuente: Encuesta de Estructura Salarial (EES), INE.

Ilustración 7. Brecha Salarial de Género en España (2008-2017)



Fuente: Encuesta de Estructura Salarial (EES), INE.

Con estas imágenes nos lanzamos al proyecto de estimar un modelo con el que obtener las ecuaciones de salarios femenina y masculina, que a su vez nos permitan conocer el salario previsto a percibir por cada individuo de acuerdo con sus características personales y profesionales.

3. Objetivos e hipótesis.

El objetivo principal de este trabajo es obtener un indicador indirecto de discriminación salarial entre hombres y mujeres en el mercado laboral español a partir de dos condicionantes:

1. El capital humano.
2. El capital cultural y social, donde entra en juego el sexo, la inserción ocupacional y la jornada laboral (el problema de la conciliación familiar).

El salario es la forma más clara y objetiva de mostrar las diferencias básicas de ingresos entre hombres y mujeres. Por este motivo no se considerarán otras fuentes de ingresos que no provengan del trabajo y además sólo se considerarán a aquellos individuos que hayan percibido un salario neto anual mayor que cero[2].

Una de las preguntas que se pueden plantear es si ¿a medida que avanza la estructura salarial la diferencia sexual tiende a aumentar?

Dado que el objetivo de este trabajo es obtener un indicador de discriminación salarial indirecta, que es parte de la brecha salarial de género, en primer lugar, se ofrecerán una serie de datos descriptivos referentes a la brecha salarial en España y, en segundo lugar, se utilizarán diferentes metodologías para estimar la ecuación de salarios para cada uno de los géneros.

Se parte de que un buen método para estimar de manera indirecta la discriminación es a través de la descomposición de Oaxaca-Blinder[3], con ella se puede averiguar en qué medida la desigualdad salarial en media viene explicada, por un lado, por las diferencias en las características profesionales de los trabajadores y las trabajadoras; y por otro, en qué medida dicha desigualdad viene explicada por otros factores (como la discriminación contra la mujer).

Para este análisis utilizaremos la Encuesta de Condiciones de Vida (ECV), elaborada por el INE. En particular proporciona información de la situación económica y los ingresos de los hogares privados, distinguiendo ente las diferentes fuentes de ingresos (rentas del trabajo, rentas del capital, pensiones, etc.) y, además, proporciona información de las principales características sociolaborales de cada miembro del hogar (educación, actividad, etc.) es una buena fuente de datos para realizar este estudio planteado.

Quiero destacar que con este trabajo se hace una innovación en el estudio de los salarios en la rama de la economía. Hasta ahora las ecuaciones de salarios siempre se han obtenido por métodos lineales, la regresión lineal, pero en este trabajo nos enfrentamos

[2] Esto crea un problema de selección muestral, pero este problema no se ha querido abordar en este estudio, dado que pretende ser un primer intento de aplicación de nueva metodología en la estimación de ecuaciones de salarios.

[3] La metodología de Oaxaca-Blinder ha sido la más ampliamente utilizada para el análisis de la desigualdad salarial en media entre hombres y mujeres. Véase Oaxaca (1973), Blinder (1973), Newmark (1988), Oaxaca y Ransom (1994), Neuman y Oaxaca (2004), y Ministerio de Sanidad, Servicios Sociales e Igualdad (20129.

al uso de algoritmos de machine learning para estimar sendas ecuaciones de salarios, como son las redes neuronales, métodos de árbol (bagging y random forest) y gradient boosting. Con ello hemos querido hacer un acercamiento de estos nuevos métodos de análisis de datos a este campo con gran controversia social e importancia económica en estos días.

4. Metodología y marco teórico del estudio.

La metodología llevada a cabo en el trabajo se basará en el esquema de la metodología SEMMA, compuesta por cinco fases:

1. Sample (muestra). En esta fase seleccionamos la muestra representativa de los datos de la ECV que nos permitirá estudiar y desarrollar los posteriores procedimientos en búsqueda de estimar los salarios.
2. Explore (exploración). En esta fase analizaremos, entenderemos y estudiaremos de forma detenida todas las variables disponibles. Con una buena documentación y consulta bibliográfica se definirán aquellas variables que anteriores estudios han considerado de interés y repercusión para la estimación de las ecuaciones de salarios que posteriormente nos permitirán estimar la brecha salarial de género. Es decir, aquellas variables que ayuden a explicar cómo varían los salarios. Para ello tendremos que definir e identificar la variable objetivo. Después nos centraremos en estudiar que variables input pueden aportar más valor y ser útiles para conseguir nuestro objetivo. Para ello haremos uso del proceso de depuración de datos. En este punto modificaremos los posibles errores, se eliminarán datos atípicos y anómalos, el tratamiento de datos missing, etc. Además, se va a realizar un análisis univariante y bivariante para ver el aporte de las variables y si existiera la necesidad de transformarlas para mejorarlas o la posibilidad de creación de nuevas variables a partir de las que ya tenemos. Esta es una de las fases más importantes del trabajo, pues una buena exploración nos permitirá una buena modelización. Esta parte se va a llevar a cabo en SAS Miner.
3. Modify (modificación). En esta fase realizaremos las modificaciones necesarias previamente identificadas para conseguir una mejor relación con la variable objetivo y de este modo obtener una mejora en la relación entre variable dependiente e input, lo que ayudará a construir un buen modelo (como la reagrupación de categorías). Se evaluará con estadísticos si tienen o no aporte cada una de las variables para iniciar la modelización.
4. Model (modelización). En este punto construiremos los distintos modelos e identificaremos a aquel que sea mejor sobre el resto, es decir, que consiga plasmar de forma óptima la relación entre la variable objetivo y el resto de las variables input para determinar de la forma más fiable el salario a percibir por cada individuo. Las diferentes alternativas de modelización que emplearemos son: regresión lineal, redes neuronales, bagging, árboles de regresión, random forest y gradient boosting.
5. Assess (evaluación). Una vez que obtengamos los modelos, los evaluaremos como de bien o mal predicen los salarios con los datos test reservados del conjunto de datos, buscando cometer el menor error posible.

El objetivo de este trabajo es cuantificar y conocer las causas de la desigualdad existente entre el salario percibido por los hombres y las mujeres en el mercado de trabajo de España. Para alcanzar estos objetivos llevaremos a cabo un análisis econométrico en el que utilizaremos la metodología de descomposición de Oaxaca-Blinder, siguiendo a Escot,L. (2018), para averiguar en qué medida la desigualdad salarial en media viene explicada por las diferencias en las características profesionales de las trabajadoras y trabajadores, y en qué medida dicha desigualdad viene explicada por otros factores como la discriminación salarial pura (por razón de sexo).

En todos los métodos pretendemos seguir un mismo proceso a la hora de la estimación de la ecuación de salarios:

1. Una partición de datos Training-Test, reservando el 70% de los datos para el entrenamiento del modelo para cada uno de los sexos y el 30% restante para la parte de validación, datos test. De este modo podremos aplicar validación cruzada repetida para verificar que los modelos se ajustan bien a los datos.
2. Estimar una ecuación de salarios para los hombres y otra para las mujeres que incluya las características que mejor definen la evolución del salario neto anual para cada uno de ellos.
3. Una vez que nuestros modelos estén validados daremos paso al punto clave[4] de nuestro objeto de estudio: ¿Qué salario debería de percibir una mujer si sus características fuesen evaluadas con la ecuación de salarios masculina? ¿Y, qué salario debería percibir un hombre si sus características fuesen evaluadas con la ecuación de salarios femenina?
4. La diferencia entre lo que percibiría una mujer evaluada con la ecuación de salarios masculina y lo que realmente percibe, en relación con el salario percibido por ésta es lo que entendemos como índice de discriminación salarial.

4.1 Discriminación Salarial. Descomposición de Oaxaca-Blinder.

El análisis de la información estadística realizado en el apartado 2, Justificación del proyecto., referente a la situación de los hombres y mujeres en el mercado laboral de España, pone de manifiesto la existencia de grandes diferencias de género entre sus salarios medios, su distribución por sectores y ocupaciones, etc. Esas diferencias en media es lo que se conoce como brecha salarial.

Siguiendo a Escot,L. (2018). El objetivo que nos planteamos es el de intentar cuantificar la discriminación salarial en media entre los hombres y las mujeres en el mercado de trabajo de España. Siguiendo a Heckman (1998), podemos definir la discriminación salarial en contra de la mujer como aquella situación en la que una mujer es tratada de diferente forma que un hombre en cuanto a su remuneración como consecuencia de su sexo, siempre y cuando no existan causas objetivas que determinen que el sexo del trabajador o trabajadora ejerza ningún tipo de efecto directo sobre su productividad. En

[4] Esto consistiría en hacer una prueba test, cómo introducir un nuevo individuo para cada ecuación de salarios y estudiar cómo se ajusta al modelo. La variable sexo es la de la ecuación con la que estaríamos trabajando, es decir, manteniendo la variable sexo constante, si varían el resto de las variables ¿qué salario se debería de percibir?

consecuencia, para poder hablar de discriminación salarial de género en el mercado de trabajo, será necesario identificar y cuantificar previamente la existencia de diferencias salariales entre hombres y mujeres con idénticos (o similares) niveles de productividad. Una vez identificada y cuantificada la existencia de discriminación, podrá realizarse la agregación de la experiencia discriminatoria individual y así obtener una medida de discriminación salarial de género en el mercado de trabajo de España.

Es difícil disponer de información sobre el salario de dos trabajadores (hombre y mujer) con idéntica productividad que permitan cuantificar la posible existencia de discriminación salarial en media. El principal problema es que la productividad de los trabajadores no es una variable directamente observable. En consecuencia, los estudios sobre la discriminación salarial proponen estimar de alguna forma dicha productividad, utilizando indicadores indirectos de los factores determinantes de la productividad de un trabajador.

La metodología que usaremos en nuestro estudio fue propuesta inicialmente por Oaxaca (1973) y Blinder (1974) y se basará en la estimación de una ecuación de salarios, es decir, en la estimación del salario recibido por un trabajador como una función de sus características.

El análisis empírico de la determinación del salario recibido por un trabajador debe realizarse atendiendo a tres tipos de factores[5]:

- factores de oferta (características /capital humano del trabajador: nivel educativo, formación, experiencia profesional, etc.)
- factores de demanda (tipología de los demandantes de factor de trabajo – tamaño de la empresa, sector de actividad, etc. – y de la relación laboral – tipo de contrato, jornada laboral, puesto y responsabilidad asumida dentro de la estructura organizativa en la empresa, posibles riesgos de accidentes laborales, desplazamientos, y otro tipo de condiciones laborales -.)
- y otros factores socioeconómicos (entre los que se encontraría la discriminación).

Los factores de oferta hacen referencia a la Teoría del Capital Humano, propuesta por Becker (1964), según la cual la productividad de un trabajador está estrechamente relacionada con su capital humano, con su capacidad o competencia en el trabajo. Los factores de demanda hacen referencia a la relación entre la remuneración salarial y las condiciones laborales (calidad del empleo), así como al tipo de trabajo a realizar que, en parte, según la teoría de los salarios hedónicos, puede determinar, en igualdad de condiciones/capacidad para el trabajo, la existencia de diferencias salariales basadas en la compensación monetaria al trabajador por aspectos tales como la peligrosidad de su empleo. Los factores socioeconómicos recogen todo tipo de variables que puedan estar explicando las diferencias salariales (estado civil, hijos a cargo, etc. del trabajador, así como la existencia en la economía de discriminación por razones de género o raza).

Con la estimación de la ecuación de salarios será posible averiguar en qué medida las características medias de un trabajador contribuyen a explicar su salario, es decir, cómo

[5] Véase Cahuc y Zylberberg (2004, cap.5) para un análisis de los factores determinantes de los salarios y las diferencias salariales.

se remunera o retribuye en media cada una de las características de un trabajador. En un entorno ausente de cualquier discriminación de género, las diferencias salariales en media entre hombres y mujeres deberían estar explicadas por las diferencias en el capital humano y los factores de demanda del trabajador medio y trabajadora media. De esta forma, podría hablarse de discriminación salarial pura cuando dos trabajadores con idéntico capital humano y con los mismos factores de demanda obtienen distinto salario.

La metodología de Oaxaca-Blinder estima y cuantifica el grado de discriminación salarial en media entre hombres y mujeres como un residuo, esto es, como la parte de la diferencia salarial entre hombres y mujeres que no puede atribuirse a diferencias en media de sus factores de oferta y de demanda.

Analíticamente, esta metodología trataría de estimar sendas ecuaciones de salarios para hombres y mujeres:

$$W_H = f^H(X_1 \dots X_K) \quad (1)$$

$$W_M = f^M(X_1 \dots X_K) \quad (2)$$

Tenemos que (1) es la Ecuación de salarios para la submuestra de hombres y (2) es la Ecuación de salarios para la submuestra de mujeres; donde *w* indica el salario recibido; f es la función usada para estimar la ecuación de salarios (puede ser: regresión lineal, red neuronal, árbol, bagging, random forest o gradient boosting), las X_n son los factores o características de cada trabajador. A partir de la estimación de las ecuaciones de salarios *(1*) y *(2*) y entendiendo que la situación de no discriminación es aquella en la que se encuentran los hombres[6], podríamos estimar cuál sería el salario de las mujeres en ausencia de discriminación w_M*, esto es, el salario que obtendrían si sus características estuviesen remuneradas igual que las de los hombres:

$$W_{M^*} = f^H(X_1 \dots X_K) \quad (3)$$

A partir de esta metodología puede definirse un indicador de discriminación relativa:

$$\%\ incremento\ salarial\ a\ recibir\ las\ mujeres = \frac{W_{M^*} - W_M}{W_M} \quad (4)$$

Este indicador recoge el porcentaje en que debería aumentar el salario de las mujeres W_M para que se igualase al salario equivalente sin discriminación W_{M^*}. Es decir, muestra el porcentaje en que debería aumentar el salario de las mujeres para que se igualase con el salario que le correspondería percibir en caso de no existir discriminación, esto es, con el que les correspondería recibir si sus características estuviesen remuneradas de igual forma que se remunera a los hombres. Este indicador permite captar la dimensión e importancia de la discriminación salarial, por lo que se enfatizará más en los resultados obtenidos con este indicador.

[6] Neumark (1988) y Oaxaca – Ransom (1994) proponen como alternativa considerar que la situación de no discriminación contra la mujer sería una situación no observable intermedia entre la que tienen los hombres y las mujeres.

Las mayores dificultades a la hora de estimar este índice de discriminación se encuentran en la estimación de las ecuaciones de salarios *(1*) y *(2*). Las distintas aportaciones de la literatura han ido señalando la existencia de distintos problemas que pueden conducir a estimaciones sesgadas de dichas ecuaciones. Una de ellas hace referencia a los problemas de selección muestral[7], que hace referencia a los problemas a la hora de obtener la muestra de datos con los que estimar las ecuaciones de salarios *(1*) y *(2*)[8]. En concreto, este problema aparece cuando no se puede observar la variable dependiente salario para todos los individuos y dicha falta de observación no es aleatorio, sino que depende de la decisión previa tomada por cada individuo sobre trabajar o no trabajar – sólo es posible observar el salario de los individuos que previamente decidieron incorporarse al mercado de trabajo[9].

En nuestro caso de estudio hemos seleccionado a aquellos individuos que han tomado la decisión de trabajar, están ocupados, y perciben un salario. Dejando a un lado a aquellos individuos que aun queriendo trabajar no perciben un salario por estar en situación de desempleo.

4.2 Ecuaciones de salarios. Métodos utilizados.

La Metodología de Oaxaca-Blinder para calcular la discriminación indirecta requiere ecuaciones de salarios de varones (*1*) y mujeres (*2*), estas ecuaciones de salarios las obtendremos usando diferentes métodos de Machine Learning. A continuación, explicaremos en que consiste cada una de las diferentes técnicas de modelización usadas.

4.2.1 Regresión Lineal.

La regresión lineal es una técnica de modelización estadística que estudia la relación entre la variable objetivo de tipo intervalo y las variables independientes.

Para analizar las diferencias salariales de género se usarán regresiones por MCO. Con ellas se podrá estimar la beta de cada uno de nuestros parámetros. Para un primer intento se modelarán del tipo:

$$\mathrm{Y} = (\beta_0 + \beta_i X_i + \beta_2 X_{ii} + \cdots + \epsilon) \qquad (5)$$

Donde Y es el ingreso salarial neto mensual. Será la variable dependiente de nuestro modelo. Recordando que el salario es todo ingreso derivado del trabajo. Las β_{ii} miden el efecto de la experiencia, nivel de estudios, sexo, … manteniendo todo lo demás constante. Y ϵ es el componente residual.

[7] Para un análisis de los problemas de selección muestral en la cuantificación de la discriminación salarial véase Neuman y Oaxaca (2003) y Hernández y Méndez (2005).

[8] Otro problema que podría sesgar la estimación de los indicadores de discriminación sería la existencia de efectos heterogéneos entre mujeres y hombres no observables que afectasen tanto al salario recibido como a las propias características de cada trabajadora o trabajador. Para poder corregir este problema habría que recurrir a la estimación de salarios a través de metodologías de Datos de Panel.

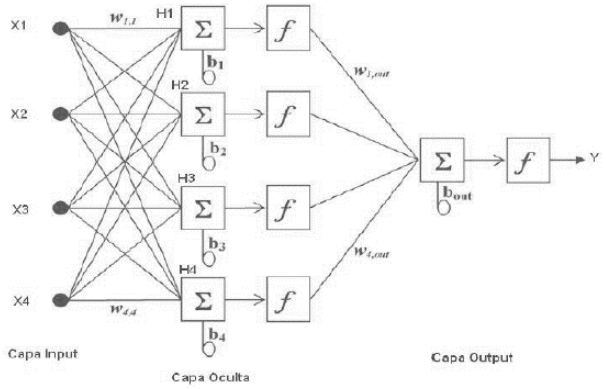
[9] Si la falta de datos sobre el salario fuese puramente aleatoria no existiría este problema, y la estimación de la ecuación de salarios sería consistente.

Cada una de las X_i representará a cada una de las diferentes variables independientes que introduzcamos en nuestro modelo. Con certeza se introducirán la edad, años de educación, experiencia (o antigüedad) y tipo de jornada.

4.2.2 Redes Neuronales

La red neuronal es un paradigma de aprendizaje y procesamiento automático (Machine Learning) inspirado en el funcionamiento del sistema nervioso humano. Una red neuronal está compuesta por un conjunto de neuronas (nodos) interconectadas entre sí mediante enlaces y organizados en grupos llamados capas. Por lo tanto, son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico.

Ilustración 8. Esquema de Red Neuronal.



Fuente: Apuntes Machine Learning (Portela,2019)

Como se puede ver en la Ilustración 8 la capa input es aquella que está formada por las variables que introducimos para modelizar (x_i), se conecta con la capa oculta (H_j) mediante la función de combinación que asigna los pesos (w_{ij}), los cuales representan la interacción entre los nodos de las capas y a cada una de estas combinaciones se les asocia un sesgo (b_{ij}). Se agregan los resultados parciales y mediante una función de activación (*f*) se calcula la salida. Esta salida es a su vez es entrada de la neurona a la que precede. Por último, se aplica la función de activación de la combinación obtenida de la capa oculta a la capa output, obteniéndose finalmente Y. La unión de todas estas neuronas interconectadas es lo que compone la red neuronal artificial.

Este proceso puede ser más o menos complejo en función del número de variables de entrada y capas ocultas. A mayor número de variables y capas ocultas la red se vuelve más compleja y el número de parámetros se incrementa. Es necesario controlar el grado de complejidad de la red de acuerdo con nuestro conjunto de datos.

Las redes neuronales necesitan entrenamiento para buscar el valor de los pesos (w_{ij}) que mejor ajustan la variable dependiente a raíz de las variables independientes. Esto hace que sea necesario una gran cantidad de datos para poder entrenar a la red, para que gane experiencia y sea robusta para poder encontrar la combinación de parámetros que mejor clasificará. Para ello, las redes neuronales necesitan especial atención a la hora de configurar una buena parametrización de los siguientes componentes:

- Número de nodos.
- Función de activación.
- Algoritmo de optimización.
- Número de iteraciones máximas, realizando pruebas de early stopping ya que en ocasiones permite evitar el sobreajuste.

4.2.3 Métodos basados en árboles.

Los árboles de regresión presentan una serie de ventajas y desventajas. Entre sus ventajas destaca la adaptabilidad a la forma funcional entre la variable objetivo y las variables predictoras, el tratamiento automático de los valores missing, el tratamiento automático de categorías poco representadas, la detección automática de regiones y puntos de corte, y resultados a menudo fáciles de comprender. Entre sus desventajas tenemos la poca capacidad predictiva y gran varianza, sensibilidad a cambios en los datos, inestabilidad y poca robustez y la falta de suavidad lo que a veces redunda en mayor error promedio de predicción en regresión.

Las desventajas de los árboles no han podido ser solucionadas mejorando las funciones de error o algoritmos de construcción, pero sí combinando el resultado de muchos árboles. Por ello se ha decidido aplicar además de los árboles de regresión, las técnicas de Bagging, Random Forest y Gradient Boosting.

4.2.3.1 Bagging

Este método parte del conjunto de datos de tamaño N del cual se seleccionan N o n>N observaciones con reemplazamiento (o sin él) de los datos originales y se aplica un árbol, del cual se pueden obtener las predicciones para todas las observaciones originales (N). Este proceso requiere que se repita al menos un número m de veces para poder promediar las m predicciones obtenidas.

4.2.3.2 Random Forest

Random forest es un método derivado del bagging, cuya característica principal consiste en incorporar aleatoriedad en las variables utilizadas para segmentar cada nodo del árbol. El procedimiento es igual que el de bagging pero con el matiz de que cada vez que se abre un nodo seleccionamos p variables de las k originales y de esas p elegidas, se escoge la mejor para llevar a cabo la partición en ese nodo.

4.2.3.3 Gradient Boosting

La técnica de Gradient Boosting (Friedman, 2001) se basa en ir actualizando las predicciones en la dirección de decrecimiento dada por el negativo del gradiente, de la función de error $L\ (y_i, f(x_i))$ (paso (a)) es la función de predicción de y_i basada en los valores x_i.

Algoritmo de árbol de Gradient Boosting.

1. Inicialmente

$$f_0(x) = \arg min_{\Upsilon} \sum_{i=1}^{N} L\,(y_i, \Upsilon) \quad (6)$$

2. Para m = 1 hasta M*:*
   a. Para i = 1, 2, … N calcula

$$r_{im} = -[\frac{\partial\,(y_i, f(x_i))}{\partial f(x_i)}]_{f=f_{mi}} \quad (7)$$

   b. Ajustar un árbol de regresión a los objetivos r_{im}dando las regiones terminales $R_{im}, j = 1,2, \dots, J_m$.
   c. Para $j = 1,2, \dots, J_m$ calcula

$$\Upsilon_{jm} = \arg min_{\Upsilon} \sum_{x_i \epsilon R_{jm}} L\,(y_i, f_{m-1}(x_i) + \Upsilon). \quad (8)$$

   d. Actualizar:

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \Upsilon_{jm} I\,(x \,\epsilon\, R_{jm}). \quad (9)$$

3. Resultado:

$$F(x) = f_M(x). \quad (10)$$

Por lo tanto el algoritmo gradient boosting consiste en repetir la construcción de árboles de regresión, modificando ligeramente las predicciones iniciales cada vez, intentando de esta manera ir minimizando los residuos en la dirección de decrecimiento.

Al plantear diferentes árboles cada vez, el proceso va ajustando las predicciones cada vez más a los datos, y de alguna forma unos árboles corrigen a otros con lo cual la flexobilidad y adapatación del método mejora respecto a la construcción de un único árbol.

Entre las ventajas del gradient boosting tenemos la invariabilidad frente a transformaciones monótonas, un buen tratamiento de missings y variables categóricas (universalidad), es muy fácil de implementar pues necesita pocos parámetros a monitorizar, tiene gran eficacia predictiva, es robusto respecto a variables irrelevantes y respecto a la colinealidad ya que detecta interacciones ocultas. Entre las desventajas destaca que para datos relativamente sencillos no tiene nada nuevo que aportar y puede ser preferible modelos más sencillos como las regresiones.

4.2.3.4 Parametrizaciones de las técnicas de árbol.

En los árboles de regresión podemos modificar los siguientes parámetros:

- Número de hojas final o profundidad del árbol.
- Número de divisiones máximas en cada nodo. Será de 2, pues trabajamos con árboles binarios.

- P-valor: determina las divisiones en cada nodo; más alto, más estricto frente a la realización de subdivisiones y más sencillos son los árboles.
- El número de observaciones mínimo que debe de haber en una rama/nodo. Si se establece un número más grande permite evitar sobreajuste (menos varianza) a cambio de algo más de sesgo, frente a reducir el número y ajustar mejor a los datos (menor sesgo, algo más de varianza en los modelos).

Para bagging además podemos modificar:

- Número de iteraciones m a promediar.
- Tamaño de la muestra n vs N para la que realizaremos los árboles.

Para random forest, además, exploraremos el número óptimo de variables p a muestrear en cada nodo.

Para gradient boosting debemos tener en cuenta:

• La constante de regularización que refleja en cuanto se modifica el error cada vez. Cuanto más elevada sea, más rápido converge el algoritmo, por lo que debemos encontrar un valor óptimo para evitar ser ni demasiado bruscos ni demasiado lentos.

• Necesidad de early stopping (parada anticipada del algoritmo para evitar sobre ajuste).

4.3 Concepto de Validación Cruzada Repetida.

Para poder evaluar la bondad de ajuste de cada uno de los modelos se usará esta técnica que reduce la dependencia entre la partición de datos de la muestra seleccionada para el entrenamiento del modelo y la partición de datos empleada para su posterior validación.

Este método consiste en dividir los datos aleatoriamente en *k* grupos, dejando un conjunto *j* a parte y construyendo el modelo con el resto de los grupos (*k – j).* Por último, con el conjunto *j* se estima el error del modelo. Este proceso para que sea más eficaz se repite varias veces (5 en nuestro caso) con variación de semillas. Además, para este trabajo hemos prefijado la creación de 4 grupos.

5. Bases de datos utilizadas.

El organismo oficial del que nos hemos nutrido y que elabora con carácter periódico estadísticas de empleo y hogares y que, por tanto, nos permite realizar análisis de los mercados de trabajo en España es el Instituto Nacional de Estadística (INE).

El INE es una fuente fiable que proporciona datos sobre la situación laboral y familiar de los hombres y mujeres en España para poder compararla entre ambos géneros, y realizar un análisis conjunto en el período de estudio (2018). Asimismo, también podemos tomar como referencia la situación general del mercado laboral en nuestro país y analizar la posición de la mujer dentro de éste. Esto nos ha permitido ver cómo han evolucionado los distintos indicadores mencionados en la sección 2 con el paso del tiempo.

5.1 Características y limitaciones de las fuentes.

La decisión de tomar la ECV como fuente de referencia para este trabajo creemos que es la más acertada por varios motivos, pero principalmente porque es la más completa a nivel general y de manera particular, para tratar los temas que nos ocupan. Además de estar homogeneizada de acuerdo con los protocolos de Eurostat, es una de las principales fuentes de información estadística sobre el mercado de trabajo, hogares y condiciones de vida en España para estudiar discriminación salarial entre hombres y mujeres, con una información de la mayor actualidad posible.

En un primer momento se consideró usar la Encuesta de Estructura Salarial, pero sólo se dispone de datos hasta 2014, dado que se trata de una encuesta de carácter cuatrienal. Aunque la Encuesta de Estructura Salarial (INE) permite conocer los salarios de cada individuo nos proporciona poca información sobre las características de los trabajadores (por ejemplo, variables familiares como el estado civil o el número de hijos) que si requerimos para la estimación de las ecuaciones de salarios. No obstante, la EES nos ha resultado útil para obtener una vista general del mercado laboral junto con la Encuesta de Población Activa en anteriores secciones.

5.2 La Encuesta de Condiciones de Vida (ECV).

La Encuesta de condiciones de vida (ECV)[10] [11] se realiza desde 2004. Es una operación estadística anual dirigida a hogares que se realiza en todos los países de Unión Europea. Su principal objetivo es proporcionar información sobre la renta, el nivel y composición de la pobreza y la exclusión social en España y permitir la realización de comparaciones con otros países de la Unión Europea.

Para ello recoge los ingresos del año natural anterior a la entrevista. Además, se recogen otras muchas preguntas sobre condiciones de vida que se refieren al momento de la entrevista. Por lo tanto, las variables sobre ingresos de la encuesta de 2018 se refieren al año 2017 mientras que el resto de las preguntas al año 2018.

Por lo tanto, su objetivo es disponer de una fuente de referencia sobre estadísticas comparativas de la distribución de ingresos y la exclusión social. El último dato disponible a fecha de 27 de junio de 2019 nos dice que la renta media por hogar en 2018 es de 28.417€ y que existe un 21,6% de población en riesgo de pobreza residente en España.

5.2.1 Características de la ECV 2018.

Es una encuesta anual. El período de recogida de los datos fue el tercer cuatrimestre de 2018. La muestra efectiva (tamaño muestral) está formada por unas 34.000 personas. En cuanto al tipo de muestreo, la ECV es una encuesta panel en la que las personas entrevistadas colaboran cuatro años seguidos. Se trata por tato de un muestreo

[10] La ECV se trata de una encuesta con criterios armonizados para todos los países de la Unión Europea. Está respaldada por el Reglamento (CE) Nº 1177/2003 del Parlamento Europeo y del Consejo de 16 de junio de 2003 relativo a las estadísticas comunitarias sobre la renta y las condiciones de vida. La realización de ésta permite poner a disposición de la Comisión Europea un instrumento estadístico de primer orden para el estudio de la pobreza y desigualdad, el seguimiento de la cohesión social en el territorio de su ámbito, el estudio de las necesidades de la población y del impacto de las políticas sociales y económicas sobre los hogares y las personas, así como para el diseño de nuevas políticas. (INE, 2019).

[11] La ECV 2018 ha sido realizada por el Instituto Nacional de Estadística (INE) en colaboración con el Instituto de Estadística de Cataluña (IDESCAT).

bietápico estratificado. Las unidades de primera etapa son las secciones censales y las de segunda etapa son las viviendas familiares.

La ECV nos ayuda a obtener información comparable y armonizada sobre diferentes aspectos del nivel y condiciones de vida y de la cohesión social. Entre ellos podemos destacar los siguientes aspectos:

- **Ingresos de los hogares privados. Situación económica.**

La ECV proporciona información sobre los ingresos, su distribución en función de ciertas características básicas, en particular: distribución del nivel de ingresos según personas y hogares; distribución de sus componentes, según personas y hogares; causas de las desigualdades en los ingresos y evolución de éstas en el tiempo.

- **Igualdad de trato del hombre y de la mujer.**

Se puede obtener información de la evolución de la familia monoparental según el número de personas a cargo (menores y otros); evolución de las tasas de actividad femenina y de las diferencias salariales; situación de la mujer en el sistema de protección social, según los indicadores de resultados relativos a las prestaciones medias por sexo, así como otros indicadores demográficos y socioeconómicos.

- **Empleo y actividad. Cuidado de niños.**

La ECV permite observar y analizar la evolución del mercado de trabajo. Recoge los movimientos a corto plazo del empleo con lo que se puede obtener datos sobre: los diversos tipos de desempleo (de corta duración, de larga duración y empleo a tiempo parcial), causas y características del empleo a tiempo parcial, causas de la rotación entre empleo y desempleo.

En lo relativo a la actividad se puede estudiar la tasa de actividad masculina y femenina, que ya hemos representado con datos de la Encuesta de Población Activa (INE) en el **¡Error! No se encuentra el origen de la referencia.** en la página 1.

Y en cuanto al cuidado de niños, se recoge el número de horas semanales que son cuidados en centros o por personas que no son sus padres, con lo que se podría analizar su vínculo con la actividad de la madre o el tipo de hogar, por ejemplo.

- **Nivel de formación, salud y efectos de ambos sobre la condición socioeconómica.**

La ECV nos permite obtener información del nivel general de formación de la población; la relación existente entre el nivel formativo y la situación socioeconómica por sexo, edad, ocupación y situación profesional; y la evolución de los ingresos según el nivel de formación.

6. Software y máquina utilizados.

Para el desarrollo de este trabajo se ha usado como herramientas de análisis estadístico R® y *SAS® software*. En R se ha usado las librerías *Caret, Randomforest, Nnet* y *Glmnet* para la modelización. Y en SAS, *SAS Miner Workstation 14.1* y *SAS Base 9.4.*

La máquina empleada ha sido un HP Notebook, Windows 10 Home, 1TB de almacenamiento, 8GB RAM, Procesador AMD A8-7410 APU.

7. Análisis descriptivo de los datos.

La ECV cuenta con una gran multitud de variables, pero sólo hemos seleccionado aquellas variables que según diferentes fuentes bibliográficas consultadas son las más interesantes para tener en cuenta en nuestro estudio. El listado completo con el detalle de todas las variables seleccionadas como las creadas se puede consultar en el Anexo 2. Variables utilizadas para estimar la ecuación de salarios.

7.1 Análisis descriptivo individual de cada una de las variables.

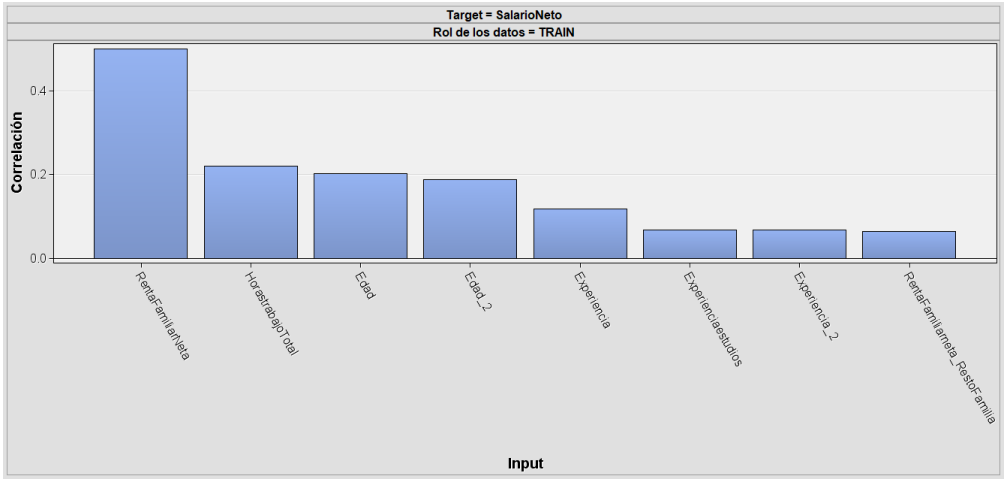
Antes de dar paso a la elaboración de nuestros modelos de estimación de ecuaciones de salarios debemos explorar cada una de las variables consideradas. En total contamos con 56 variables, de las cuales 47 variables son categóricas (44 dicotómicas y 3 nominales), 8 de intervalo y una variable objetivo de tipo intervalo que dan lugar a un conjunto de datos formado por 11.765 observaciones, 6.207 observaciones para hombres y 5.558 para mujeres, tras el muestreo previo filtrando por la variable *Mujer* (1 sí es mujer, 0 si es hombre).

Para las variables de tipo intervalo hemos obtenido los estadísticos descriptivos de cada una de ellas (mediana, missings, mínimo, máximo, media, desviación estándar y asimetría) con SAS Miner, la salida con detalle de los estadísticos se puede consultar en el *Anexo 2. Variables utilizadas para estimar la ecuación de salarios*., en las Tabla 36 y Tabla 37.

De igual forma hemos obtenido el número de ausentes, los niveles por cada categoría, la moda y el porcentaje de la moda para las variables categóricas. Estas se pueden consultar en la Tabla 38 del mismo anexo.

7.1 Análisis descriptivo bivariante entre cada una de las variables input y la variable dependiente salario neto.

Ilustración 9. Correlación de Pearson entre variables input de intervalo y objetivo SalarioNeto, muestra mujeres.



El coeficiente de correlación de Pearson es una medida lineal entre dos variables aleatorias cuantitativas y continuas. A diferencia de la covarianza, este coeficiente es independiente de la escala de medida de las variables. La fórmula empleada para su cálculo se puede expresar como:

Suponiendo que estamos estudiando dos variables aleatorias X e Y sobre una población, el coeficiente de correlación de Pearson se puede representar como $\rho_{X,Y}$, siendo la expresión para su cálculo:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \quad (11)$$

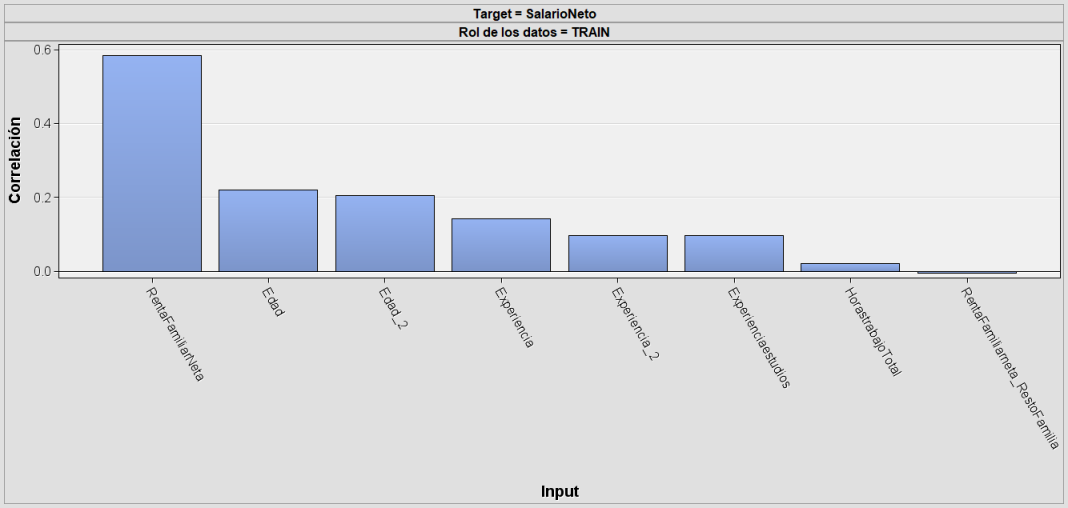
Donde:

- σ_{XY} es la covarianza de (X, Y)
- σ_X es la desviación estándar de la variable X
- σ_Y es la desviación estándar de la variable Y

Este índice puede tomar valores entre el intervalo [-1, 1], indicando el signo el sentido de la relación. Para valores de $\rho = 1$, existe una correlación perfecta positiva. Para valores de $\rho = -1$ indica una correlación perfecta negativa[12]. Para ambos valores existe una dependencia total entre las dos variables. Para el primer caso tendríamos una relación directa (cuando una aumenta la otra lo hace en proporción constante) y para el segundo caso inversa (cuando una aumenta la otra disminuye en proporción constante).

Como se puede ver en la Ilustración 9 únicamente se ha estimado el coeficiente de correlación de Pearson entre el salario neto mensual de los individuos entre 16 y 64 años y las variables input de intervalo horas de trabajo en total, renta familiar neta, edad, edad^2, experiencia en los estudios, experiencia, experiencia^2 y la renta familiar neta del resto de la familia.

Ilustración 10. Correlación de Pearson entre variables input de intervalo y objetivo SalarioNeto, muestra hombres.



Para la Ilustración 10 hemos representado el coeficiente de correlación de Pearson de cada una de las variables input de intervalo con la variable objetivo para los hombres. En este caso podemos ver que la renta familiar neta del resto de la familia tiene correlación negativa.

[12] Además de los dos valores mencionados existen otros valores intermedios. Si $0 < \rho < 1$ existe una correlación positiva; si $\rho = 0$ no existe relación lineal, pero no necesariamente implica que las variables son independientes pues, pueden existir relaciones no lineales entre las dos variables; y si $-1 < \rho < 0$ existe una correlación negativa.

En nuestro conjunto de datos hemos añadido las variables Experiencia^2 y Edad^2, dos transformaciones cuadráticas de las variables Experiencia y Edad, respectivamente, para estudiar cómo afecta cada una de ellas al salario neto anual.

Ilustración 12. Correlación de Pearson para mujeres.

Input	Correlación
RentaFamiliarNeta	0.50106
HorastrabajoTotal	0.22054
Edad	0.20182
Edad_2	0.18739
Experiencia	0.11768
Experienciaestudios	0.06715
Experiencia_2	0.06706
RentaFamiliarneta_RestoFamilia	0.06355

Ilustración 11. Correlación de Pearson para hombres.

Input	Correlación
RentaFamiliarNeta	0.58400
Edad	0.22058
Edad_2	0.20605
Experiencia	0.14168
Experiencia_2	0.09689
Experienciaestudios	0.09644
HorastrabajoTotal	0.02133
RentaFamiliarneta_RestoFamilia	-0.00357

Una vez calculado este estadístico podemos obtener información muy interesante para la estimación de las ecuaciones de salarios. Para ello nos centraremos en las cifras proporcionadas en la Ilustración 12 y la Ilustración 11.

1. En primer lugar, las variables más correlacionadas positivamente con el salario son la renta familiar neta, las horas de trabajo en total y la edad, para mujeres y la renta familiar neta y la edad para los hombres.
    a. Por cada euro percibido el 22,05% es explicado por las horas de trabajo desempeñadas en las mujeres, y el 2,13% en los hombres. A más horas trabajadas más salario.
    b. Por cada euro percibido el 20,18% es explicado por la edad de la mujer y el 22,06% por la edad del hombre. A más edad mayor salario.
    c. Por cada euro percibido el 50,11% es explicado por la renta neta percibida por la familia para las mujeres y el 58,4% para los hombres. Esto se puede entender por el tramo de IRPF. A más renta familiar más salario.
2. En segundo lugar, existe una correlación positiva entre el salario y la experiencia en los estudios y la experiencia laboral.
    a. Por cada euro percibido el 6,72% es explicado por la experiencia de los estudios, entendido como los años en práctica del capital humano adquirido por la educación, para las mujeres y del 9.65% para los hombres. A mayor experiencia en los estudios mayor salario. De aquí podemos deducir que el nivel educativo es una variable que discretiza muy bien el salario percibido, donde a más nivel educativo mayor salario.
    b. Por cada euro percibido el 11.77% es explicado por la experiencia laboral adquirida para las mujeres y el 14,17% para los hombres. A mayor experiencia laboral, más preparación y adaptación al cambio, versatilidad en el puesto de trabajo, lo que aporta un plus.
3. Por último, existe una correlación positiva entre la renta familiar neta del resto de la familia para las mujeres, pero negativa para los hombres. Es un arma de doble filo:
    a. Por cada euro percibido, el 6.36% viene explicado por la renta familiar neta percibida por el resto de la familia para las mujeres. A mayor renta

familiar obtenida por el resto de los integrantes de la familia, se opta por trabajar más por parte de las mujeres, dado que pueden buscar por ejemplo alternativas que le permitan la compatibilidad entre trabajo y familia.

b. Por cada euro que se deja de percibir por el salario el 0.36% viene explicado por la renta familiar neta del resto de la familia para los hombres. Los hombres tienen un incentivo, aunque sea mínimo, a trabajar menos si la renta familiar del resto de la familia es elevada.

Además de la utilización del coeficiente de Pearson para determinar que variables son las más correlacionadas con la variable objetivo hemos empleado una utilidad de SAS Miner denominada Nodo Multigráfico[13] que nos permite concluir que variables van a resultar más útiles para la estimación de las ecuaciones de salarios y el Nodo Explorador de Estadísticos, que calcula varios estadísticos para cada una de las variables. Los gráficos que justifican porque se han seleccionado como variables más importantes las enumeradas y explicadas, se encuentran en el Anexo 3. Variables más importantes para la estimación de la ecuación de salarios.

7.1.1 Variables más importantes para la estimación de la ecuación de salarios.

Una vez obtenidos los gráficos de dispersión, los coeficientes de correlación de Pearson y el gráfico de importancia de la variable para cada una de las variables input contempladas para la estimación de las ecuaciones de salarios podemos decir que las variables más útiles son:

- Actividad de la empresa. No tiene el mismo salario medio las empresas dedicadas al sector de la construcción que las dedicadas a la industria química, por ejemplo.
- Casado o conviviendo. El salario en media percibido por un individuo de la población es diferente según sea su estado civil.
- Comunidad Autónoma. En cada Comunidad Autónoma la retribución del trabajo difiere en media entre ellas.
- Edad. El salario medio difiere según el tramo de edad en el que se encuentre el trabajador. Recordamos que sólo hemos considerado a la población entre 16 y 64 años que son la población en edad de trabajar.
- Estudiante. Una persona que estudia y trabaja percibe un salario diferente al de una persona que sólo trabaja. Refleja el problema de compaginar trabajo con estudios.
- Experiencia. La experiencia es un plus en el salario, por ello existe diferencias retributivas entre una persona sin experiencia y otra que si la tiene.
- Experiencia en los estudios. Los años en lo que se lleva en práctica los conocimientos adquiridos durante la etapa de estudios junto con la experiencia que se obtiene al ponerlos en práctica supone una diferencia en el salario.
- Extranjero. Existe una diferencia media entre el salario medio percibido por una persona proveniente de otro país distinto de España y entre una persona nativa.
- Horas de trabajo en total. A mayor número de horas de trabajo mayor retribución.

[13] Para el Nodo Multigráfico de SAS Miner hemos activado la opción de "Ambos" para la generación de los gráficos de entrenamiento. De esta forma obtenemos gráficos de barras y de dispersión que nos ayudan para determinar de forma visual que variables son más interesantes y útiles.

- Limitación de salud. Tener una limitación de salud grave, leve o ninguna explica que ciertos individuos tengan mayor o menor salario.
- Máximo nivel de estudios completados. A mayor nivel de estudios completados en principio se accede a mejores puestos de trabajo con mejor remuneración, por ello el nivel medio para cada máximo nivel de estudios finalizados varía.
- Número de hijos[14]. El número de hijos promedio varía en el salario percibido.
- Ocupación. Aquellas personas que están ocupadas perciben salario a diferencia de las que no lo están.
- Renta familiar neta. El salario varía según el tramo de renta neta de las familias.
- Renta familiar neta del resto de la familia. El salario varía según el tramo de renta neta del resto de la familia.
- Salud. Tener una salud buena, regular, mala o muy mala difiere los salarios medios percibidos por la población.
- Supervisor. Las personas con un cargo de supervisor o semejante en su lugar de trabajo perciben un salario diferente de los que no lo son.
- Tamaño de la empresa. La media de salarios varía según el tamaño de la empresa, es decir, según el número de empleados tenga.
- Temporal. La temporalidad del trabajo, como un trabajo de verano o de navidad, influye en tener un salario mayor o menor que el de otro individuo con un trabajo sin carácter de temporalidad.
- Tiempo parcial. Relacionada con las horas de trabajo en total, un puesto de trabajo con menor número de horas tiene un salario inferior que el mismo a jornada completa.
- Zona donde se vive y/o trabaja. El tamaño del municipio o unidad donde se trabaja influye en el salario. No es lo mismo trabajar en una ciudad pequeña que en una metrópolis.

8. Depuración de datos

Damos paso a uno de los puntos más importantes para realizar un buen análisis de datos: la depuración de datos. La depuración de datos es el proceso por el cuál adecuamos el conjunto de datos para la fase de modelización posterior. “Limpiamos” el conjunto de datos original de manera que no haya observaciones incoherentes y para que las variables se ajusten a las especificaciones del modelo, entre otras. La depuración de datos comprende las fases de exploración y modificación del proceso SEMMA.

Como se puede ver en las tablas del Anexo 2. Variables utilizadas para estimar la ecuación de salarios. Hay varias variables que tienen datos missing. La base de datos del

[14] Los autónomos y/o sociedades cuando contratan a un trabajador tienen la obligación de descontar las retenciones del IRPF en su nómina e ingresarlas trimestralmente en hacienda en su nombre. Para ello los pagadores recaban la información necesaria para poder calcular las retenciones. Los datos personales requeridos son entre otros: situación familiar, grado de discapacidad, número de descendientes a su cargo, mayores de 65 a su cargo, familiares discapacitados a cargo, pensiones compensatorias en favor de hijos o cónyuge, pagos por adquisición de vivienda con hipoteca, duración del contrato y tipo de relación laboral. No obstante, en la campaña de renta correspondiente Hacienda hace los cálculos oportunos para determinar la renta neta finalmente obtenida.

INE es muy robusta, por lo que la ausencia de estos datos puede ser a falta de respuesta por parte de los encuestados o a que no procede la respuesta en el caso del individuo entrevistado. No obstante, antes del tratamiento de los datos missing haremos una detección previa de datos atípicos u outlier.

Un dato atípico es una observación que es numéricamente distante del resto de los datos. El problema de los datos atípicos es que tienen una gran influencia en los resultados, si los modelos no son robustos. Una vez que son detectados tenemos dos opciones: eliminar la observación por completo o poner ese valor como missing para después imputarlo si fuera necesario.

Para una buena detección de atípicos usaremos al menos dos métodos para corroborar que la variable necesita un tratamiento de atípicos, es decir, una doble confirmación. Los métodos que vamos a usar con SAS Miner® serán:

- Desviación típica. Se considerarán como atípicos aquellos datos que disten más de un número K (habitualmente entre 3 y 6) de desviaciones típicas de la media. Este método sólo es válido si las distribuciones son aproximadamente simétricas.
- Median Absolute Deviation (MAD). Se considerarán como atípicos aquellos datos que disten más de un número k (habitualmente entre 8 y 15) de MADs de la mediana. Este método es más adecuado para distribuciones asimétricas, pero no es válido cuando la mediana es igual a 0.
- Rango Intercuartílico. Se considerarán atípicas aquellas observaciones que se alejan más de 1,5 o 3 veces el rango intercuartílico del primer y el tercer cuartil. Este método se asocia a los gráficos de cajas.

Tabla 2. Métodos de detección de atípicos por MAD y Desviación Estándar para las variables de intervalo. Mujeres.

Variable	Mínimo	Máximo	Método de Filtrado
Edad	12.8629576	76.24031691	STDDEV
Edad_2	-667.182129	4859.988174	STDDEV
Experiencia	-11.2396041	58.73987454	STDDEV
Experiencia_2	-1011.47827	2408.858627	STDDEV
Experienciaestudios	-14.1697587	60.83834453	STDDEV
HorastrabajoTotal	6.56681214	65.62678268	STDDEV
RentaFamiliarNeta	-8610634.5	15553142	MADS
RentaFamiliarneta_RestoFamilia	-7270908.5	11544069	MADS
SalarioNeto	-65554.6	94910.9	MADS

Tabla 3. Métodos de detección de atípicos por Rango Intercuartílico para variables de intervalo. Mujeres.

Variable	Tipo de característica del gráfico de caja	Mínimo	Máximo
HorastrabajoTotal	Farhigh	56	90
	Farlow	2	19
RentaFamiliarNeta	Farhigh	13186540	29252100
RentaFamiliarneta_RestoFamilia	Farhigh	9704910	27396180
SalarioNeto	Farhigh	80187.2	229012.6

Tabla 4. Métodos de detección de atípicos por MAD y Desviación Estándar para las variables de intervalo. Hombres.

Variable	Mínimo	Máximo	Método de Filtrado
Edad	11.8804651	77.2637269	STDDEV
Edad_2	-735.011068	4945.81468	STDDEV
Experiencia	-10.7169242	61.3312652	STDDEV
Experiencia_2	-1035.44232	2600.89487	STDDEV
Experienciaestudios	-14.0893023	62.852634	STDDEV
HorastrabajoTotal	17.4884039	65.3377601	STDDEV
RentaFamiliarNeta	-8057400	14812860	MADS
RentaFamiliarneta_RestoFamilia	-7526570	10992190	MADS
SalarioNeto	-79045.6	116537	MADS

Tabla 5. Métodos de detección de atípicos por Rango Intercuartílico para variables de intervalo. Hombres.

Variable	Tipo de característica del gráfico de caja	Mínimo	Máximo
HorastrabajoTotal	Farhigh	41	99
	Farlow	3	39
RentaFamiliarNeta	Farhigh	12733560	25898080
RentaFamiliarneta_RestoFamilia	Farhigh	9341470	23577117
SalarioNeto	Farhigh	97106.5	398312.6

Tras aplicar los métodos de desviación estándar y MAD para todas las variables en una primera inspección hemos obtenido los límites sugeridos en la Tabla 2 para las mujeres y en la Tabla 4 para los hombres. Para verificar si es necesario aplicar límites con un nodo de código hemos aplicado el método de Rango Intercuartílico para todas las variables Con el método de rango intercuartílico hemos obtenido que sólo se aplicaría límites a cuatro variables: las horas de trabajo en total, la renta familiar neta, la renta familiar neta del resto de la familia y la variable dependiente salario neto, como se puede ver en la Tabla 3 y la Tabla 4.

Por lo tanto, sólo aplicaremos límites a las cuatro variables mencionadas, pues para ellas se ha comprobado que por al menos dos métodos requieren un tratamiento de atípicos. Los límites se aplicarán de la manera menos restrictiva, que abarque mayor rango, esto es, el menor de los mínimos y el mayor de los máximos sugeridos por cada uno de los métodos. Siempre desde un punto de vista lógico (por ejemplo: la experiencia no puede ser negativa). Para la variable *horas de trabajo en total* por rango intercuartílico hemos obtenido farhigh y farlow para el gráfico de caja, de tal manera que hemos de seleccionar el menor valor del farhigh y el mayor valor del farlow para establecer los mínimos y máximos a comparar con los obtenidos por el método de desviación típica.

En la

Tabla 6 figuran los límites que finalmente se han aplicado a las variables de las mujeres y en la Tabla 7 para los hombres.

Tabla 6. Límites aplicados a las variables de intervalo en el proceso de detección de atípicos. Mujeres.

Variable	Mínimo	Máximo
HorastrabajoTotal	6.566812139	65.6267827
RentaFamiliarNeta	0	15553142
RentaFamiliarneta_RestoFamilia	0	11544069
SalarioNeto	0	94910.9

Tabla 7.Límites aplicados a las variables de intervalo en el proceso de detección de atípicos. Hombres.

Variable	Mínimo	Máximo
HorastrabajoTotal	17.4884039	65.3377601
RentaFamiliarNeta	0	14812860
RentaFamiliarneta_RestoFamilia	0	10992190
SalarioNeto	0	116537

NOTA: Dado que para rango intercuartílico hemos obtenido un límite inferior (FARHIGH) para la variable *Horas de trabajo en total* aplicamos el límite inferior sugerido por el método de desviación estándar por ser el menos restrictivo en ambas muestras de población.

8.1 Datos faltantes o missing.

La presencia de los datos faltantes en nuestro conjunto de datos ha de ser analizada con detenimiento pues da lugar a una reducción del número de observaciones válidas para los procedimientos estadísticos posteriores. "Los datos missing o faltantes son un grave problema a la hora de realizar cualquier estudio, ya que la mayoría de los análisis en estadísticas dan por hecho que la información está completa para todas las variables del análisis" [15].

Si la presencia de missings no es aleatoria podría venir acompañada de sesgo en las respuestas y, por tanto, en los modelos (por ejemplo, encuestados que se niegan a responder preguntas).

Por tanto, analizaremos detalladamente los missings y optaremos por una de las estrategias siguientes:

- Eliminación: de variables como de observaciones. La eliminación sólo se lleva a cabo cuando la proporción de missings sea muy elevada y la pérdida de información no lo sea tanto.
- Recategorización: de los valores missing como una categoría válida. Para las variables continuas implicaría una discretización de esta.
- Imputación: sustituir los missings por valores válidos. La importancia de la imputación de los datos faltantes es porque en caso de que no se haga, la ausencia de algunas de las observaciones puede afectar de manera importante a la muestra.

Cuando existe falta de información, es necesario saber si estos datos están distribuidos aleatoriamente o si se puede identificar algunas pautas. Una correcta evaluación de los datos ausentes nos puede ayudar a evitar una reducción de la muestra o incluso tener sesgos potenciales. Para ello vamos a hacer un análisis de la proporción de missings.

- Variables. Con el nodo de DMDB obtenemos la información sobre el número de missings de cada variable.
- Observaciones. A veces existen observaciones con un gran número de missings que aportan poca información y que pueden ser eliminados. Para saber el

[15] v. d. H. GJ, «NCBI,» 11 Julio 2006. [En línea]. Available: https://www.ncbi.nlm.nih.gov/pubmed/16980151.

número de missings por observación hemos creado una variable que los cuenta para utilizar dicha información para eliminar si es el caso de las observaciones "conflictivas".

"Hay que realizar una distinción entre mecanismos y patrones de perdida de datos. Los patrones describen los datos que son observados y los que están perdidos, mientras que los mecanismos describen el proceso mediante el cual se han dado esos valores perdidos"[16].

En casi todas las referencias sobre datos perdidos se distinguen tres mecanismos (Revenga, J.M.A, 2018)[17]:

- MCAR (perdidos completamente al azar). Los datos missing de la variable que se estudie se dice que son completamente aleatorios si la presencia del dato perdido no depende del valor que tome dicha variable ni de ninguna otra variable que se tenga en el conjunto de datos. Este mecanismo es el más fácil de tratar. Por ejemplo, se pierden datos de recogida.
- MAR (perdidos al azar). Los datos se dicen que son perdidos al azar cuando la probabilidad de que aparezca un dato perdido en la variable Y no depende de dicha variable, pero si puede depender de una o más variables que estén el fichero de datos.
- MNAR (datos no perdidos aleatoriamente). En este mecanismo la probabilidad de pérdida de un dato puede depender tanto de la misma variable como de los valores observados de las demás variables.

Una vez que se ha analizado el grado de aleatoriedad y sus pautas en los datos faltantes tenemos que tomar la decisión de eliminarlos o imputarlos.

La presencia de datos missing es elevada y la pérdida de información también lo es. Tras el tratamiento de datos se ha encontrado una relación entre varias variables. Cuando se elimina los datos missing por observación perdemos información sobre aquellas personas que no están ocupadas, que están paradas, que son inactivas y que no son extranjeros. Esto se puede deber a que hemos creado variables binarias a partir de los datos originales (recategorización del conjunto de datos). Por lo que eliminar los datos missing no es una buena elección. Además, la eliminación de datos missing no es aconsejable a no ser que el número de missing sea muy pequeño.

Por lo tanto, decidimos usar como metodología para el tratamiento de los datos faltantes la imputación de estos a partir de las otras variables del fichero. Para ello disponemos de varios tipos de imputación:

- Imputación simple:
    - Se sustituyen los missings por algún estadístico de localización, como la media o la mediana para variables de intervalo, o la moda para variables categóricas.
    - Se imputan los datos faltantes aleatoriamente teniendo en cuenta la distribución de la variable.

[16] Á. Planchuelo, «Trabajo Fin de Máster,» Julio, Madrid, 2017.
[17] J. M. A. Revenga, «Datos Missing: Detección y tratamiento».

- Imputación por modelos: se sustituyen los missings por una predicción basada en otras variables del conjunto de datos.

La imputación se aplicará a través del nodo Imputar de SAS Miner®. Decidimos hacer una imputación por modelos obteniendo la predicción a través del método de Árbol. De esta forma se crea un árbol de regresión que estima el mejor valor para cada uno de los datos ausentes a partir de los demás datos de los que se dispone.

8.2 Recategorización de variables.

Antes de dar paso al proceso de selección de variables, en SAS Miner® con el nodo de Selección de Variables, hemos hecho una prueba de selección de variables con el estadístico r-cuadrado para estudiar el aporte de la variable input al output. Nos ha sugerido una reagrupación de las variables categóricas *Máximo nivel de estudios, Número de Hijos y Comunidad Autónoma*. La reagrupación final se puede ver en la Tabla 40 en el Anexo 2. Variables utilizadas para estimar la ecuación de salarios.

9. Selección de Variables.

Hasta ahora contamos con una gran cantidad de variables para la estimación de las ecuaciones de salarios, pero en el tratamiento estadístico de los datos se contempla la posibilidad de realizar una selección de variables previas a los modelos. ¿En nuestro caso es óptimo realizar una selección de variables o no? Para ello desarrollaremos un poco la teoría sobre la selección de variables. En esta sección explicaremos la necesidad o no de seleccionar variables en el modelo de regresión.

Disponemos de un conjunto grande de posibles variables explicativas, pero una posible pregunta sería saber si todas las variables deben entrar en el modelo de regresión, y en caso negativo, saber que variables deben entrar y cuáles no.

Los métodos de selección de variables se encargan de abordar el problema de construcción o selección del modelo. De manera general, si se incluyen cada vez más variables en un modelo de regresión, el ajuste de los datos mejora, aumenta la cantidad de parámetros a estimar, pero disminuye su precisión individual (mayor varianza) y por tanto la de la función de regresión estimada, se produce un sobreajuste. Por el contrario, si se incluyen menos variables de las necesarias en el modelo, las varianzas se reducen, pero los sesgos aumentan obteniéndose una mala descripción de los datos. Por otro lado, algunas variables predictoras pueden perjudicar la confiabilidad del modelo, especialmente si están correlacionadas con otras. De este modo, el objetivo de los métodos de selección de variables es buscar un modelo que se ajuste bien a los datos y que sea posible buscar un equilibrio entre bondad de ajuste y sencillez.

Los métodos para seleccionar variables que utilizaremos serán los siguientes:

- ❖ **Algoritmos**: entre los algoritmos para seleccionar variables podemos destacar los siguientes:

- **Métodos Forward**: consiste en la selección de variables hacia delante. Se parte de un modelo muy sencillo y se van agregando términos con algún criterio, hasta que

no procede añadir ningún término más, es decir, en cada etapa se introduce la variable más significativa hasta una cierta regla de parada.

- **Métodos Backward**: se basa en la eliminación de variables hacia atrás. Se parte de un modelo muy complejo, que incorpora todos los efectos que pueden influir en la respuesta, y en cada etapa se elimina la variable menos influyente, hasta que no procede suprimir ningún término más.
- **Métodos Stepwise**: engloba una serie de procedimientos de selección automática de variables significativas, basadas en la inclusión o exclusión de estas en el modelo de una manera secuencial. Es una combinación de los anteriores. Comienza como una introducción progresiva, pero en cada etapa se plantea si todas las variables introducidas deben de permanecer en el modelo.

Cuando se aplica este tipo de procedimientos tenemos que tener en cuenta cual será la condición para suprimir o incluir un término. Para ello podemos considerar dos criterios: criterios de significación del término y criterios de ajuste global.

- **Criterios de significación**: en un método **backward** se suprimirá el término que resulte menos significativo, y en un método **forward** se añadirá el término que al añadirlo al modelo resulte más significativo. Un criterio de significación puede ser la significación de cada coeficiente.
- **Criterios globales**: En vez de usar la significación de cada coeficiente, podemos basarnos en un criterio global, una medida global de cada modelo, de modo que tenga en cuenta el ajuste y el exceso de parámetros. Escogeremos el modelo cuya medida global sea mejor. Como criterios destacamos el Criterio de Información de Akaike (AIC), el Criterio de Información Bayesiano (BIC) y el Criterio Bayesiano de Schwarz (SBC). Se trata de buscar un modelo cuyo AIC o BIC o SBC sea pequeño, ya que en ese caso habría una verosimilitud muy grande y pocos parámetros.

  - **Métodos de mínimos cuadrados penalizados**: se basan en los mínimos cuadrados ordinarios, pero añadiendo una penalización en la función objetivo, para forzar que alguna componente del vector de parámetros sea cero y de esta manera conseguir la estimación de los parámetros y selección de variables conjuntamente.

Método LASSO: (Least Absolute Shrinkage and Selection Operator) introducido por Tibshirani (1996) es un método que combina contracción de algunos parámetros hacia cero y selección de variables, imponiendo una restricción o una penalización sobre los coeficientes de regresión. Se basa en limitar el número de parámetros del modelo de regresión clásico. Matemáticamente consiste en obtener los valores de β de la recta de regresión tales que:

$$\min_{\beta} \sum_{i=1}^{n} [y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)]^2$$

Sujeto a:

$$\sum_{j=1}^{m} |\beta_j| = t$$

Siempre que se asigne al parámetro t un valor inferior a la suma de los parámetros mínimo-cuadráticos, las betas se verán reducidos. Nótese que la expresión anterior se puede expresar de forma equivalente como:

$$\min_{\beta} SSE + \lambda \sum_{j=1}^{m} |\beta_j|$$

Siendo λ el parámetro de regularización o de penalización.

Para valores grandes de λ o valores pequeños de t, los coeficientes β_j se contraen hacia cero y alguno de ellos se anula, por eso se dice que LASSO produce estimación de parámetros y selección de variables simultánea.

La dificultad consiste en determinar el valor óptimo de t o λ. Lo habitual es probar con varios valores y seleccionar el que mejores resultados ofrezca (en términos de ASE – ya sea en el conjunto de datos de validación o mediante validación cruzada -, de AIC, BIC, SBC, etc.).

9.1 Aplicación práctica del proceso de selección de variables en el conjunto de datos

El proceso de selección de variables lo hemos hecho con R® y *SAS® Base®software* para comparar los resultados obtenidos en este proceso tan interesante.

Con el software de R® hemos elaborado un código[18] para llevar a cabo el proceso de selección de variables (Calviño, A.;2019) con las diferentes formas mencionadas anteriormente (hacia delante, hacia atrás y paso a paso).

En primer lugar, se ha renombrado a todas las variables para facilitar todo el estudio que se inicia desde este punto (ver Tabla 8). En segundo lugar, se ha realizado una partición *Training-Test* reservando el 70% de los datos para entrenar y el 30% restante para datos test/validación en cada uno de los conjuntos de datos antes de iniciar el proceso de selección de variables. El 30% reservado será el que se usará para validar finalmente los modelos obtenidos y para el cálculo de las predicciones de salarios.

No obstante, para el proceso de selección de variables hemos hecho otra partición de datos training- test del 70-30%. En cada iteración de los métodos se ha evaluado el AIC/SBC del modelo resultante. Para ello ha sido necesario indicar el rango de los modelos a evaluar, que va desde el menor (sólo con la constante) al mayor modelo posible.

[18] El código usado se puede consultar en el Anexo 4. Código usado en SAS® software y en R®.

Tabla 8. Renombre de Variables para modelizar.

Renombre	Nombre original	Renombre	Nombre original
X1	ZMuyPoblada	X30	Act11_Sanid
X2	ZMedioPoblada	X31	Act12_OtrosSer
X3	ZPocoPoblada	X32	SaludMuyBuena
X4	Edad	X33	SaludBuena
X5	Extranjero	X34	SaludRegular
X6	Estudiante	X35	SaludMala
X7	MaximoNivelEstudios	X36	SaludMuyMala
X8	Experienciaestudios	X37	LimitacionSaludGrave
X9	Asalariado	X38	LimitacionSaludLeve
X10	Ocu1_Direc	X39	LimitacionSaludNinguna
X11	Ocu2_Tecnicos	X40	RentaFamiliarNeta
X12	Ocu3_Tecapoyo	X41	RentaFamiliarneta_RestoFamilia
X13	Ocu4_Administrativo	X42	Edad_2
X14	Ocu5_Servicios	X43	Aut_nomo
X15	Ocu6_Agricultura	X44	Tiempo_Parcial
X16	Ocu7_Cualificado	X45	Experiencia_2
X17	Ocu8_Operador	X46	IMP_CasadooConviviendo
X18	Ocu9_NoCualif	X47	IMP_Experiencia
X19	HorastrabajoTotal	X48	IMP_Supervisor
X20	Act1_Agr	X49	IMP_Tamanoempresa_11_19
X21	Act2_Ind	X50	IMP_Tamanoempresa_1_10
X22	Act3_Constr	X51	IMP_Tamanoempresa_20_49
X23	Act4_Comer	X52	IMP_Tamanoempresa_50omas
X24	Act5_Hostel	X53	IMP_Temporal
X25	Act6_Trans	X54	REP_Comunidad_Aut_noma
X26	Act7_Finan	X55	REP_IMP_NumHijos
X27	Act8_ServEmpr	Y	SalarioNeto
X28	Act9_AAPP		
X29	Act10_Educ		

Para poder evaluar si es necesario hacer la selección de variables, previamente al proceso de selección de variables, hemos estimado el modelo de regresión lineal con todas las variables al que hemos denominado *ModeloPreliminar* y una regresión Lasso[19] con todas las variables a la que hemos denominado *Cv.lasso.* Al igual que se realizó para el resto de los modelos le pedimos a R que nos proporcione el valor del R^2 para evaluar la bondad del modelo a partir de los datos test (30% de la muestra).

En segundo lugar, hemos hecho una selección de variables sin interacciones. Que abarca desde el modelo más sencillo (usando el término constante como única variable regresora) hasta el modelo con todas las variables. Se ha hecho para el criterio AIC y SBC (BIC en R).

En tercer lugar, hemos hecho una selección de variables con interacciones. Se ha realizado aplicando los métodos Stepwise y Backward y para los criterios AIC y BIC. Pero

[19] La regresión Lasso se ha llevado a cabo en R con la librería *glmnet.* Para determinar el mejor valor del parámetro se ha usado validación cruzada, con el que se ha obtenido un gráfico en el que se representa para cada valor el ASE resultante.

después de probarlos, se obtenían resultados que no parecían relevantes, dado a las interacciones sugeridas entre las variables, por lo que se han descartado del análisis.

Para evaluar todos los modelos hemos pedido a R® que nos calcule el R^2 a partir de los datos test de la partición que se ha hecho sobre los datos train del proceso de selección.

Finalmente hemos comparado todos los modelos a partir de validación cruzada. Este método consiste en dividir el conjunto de datos en submuestras e iterativamente construir el modelo con todas las observaciones menos las de una submuestra y evaluarlo a continuación con las observaciones de dicha submuestra excluida. La comparación de modelos se lleva a cabo a partir del error cuadrático medio (ASE) que se obtiene como la SSE dividida entre el número de observaciones. Una forma de interpretar fácilmente los resultados obtenidos es construir un diagrama de cajas con todos los RMSE (o R^2, en nuestro caso) obtenidos al repetir el proceso de validación cruzada. Al representar los diagramas de caja para los distintos modelos sobre la misma escala se puede concluir qué modelo es el preferible respecto al resto.

La regresión Lasso para ambas muestras nos sugiere que la landa (o parámetro de penalización) estaría en torno a -2.9. Sabemos que el Lasso empieza con un modelo que incluye a todas las variables y de ahí va sacando las que no considera necesarias. En ambas muestras considera a todas las variables, es más, en el conjunto de las variables categóricas solo falta una de ellas por cada categoría, por lo que no nos sugiere ninguna fusión (si faltase dos o más sugiere fusión). Es prácticamente igual al *ModeloPreliminar* con la salvedad de que tiene un peor R^2.

Con la validación cruzada realizada para 20 repeticiones hemos obtenido el R^2 en media para cada uno de los modelos considerados. Además, hemos calculado el número de parámetros de cada uno de los modelos para determinar qué tan bueno es el modelo.

El criterio de selección del mejor modelo se basa en aquel que requiera menor número de parámetros y tenga un mayor R^2. Y que además sea lo suficiente representativo y contenga a las variables regresoras más importantes para la estimación de una buena ecuación de salarios.

En SAS Base usando la *macro randomselect* (Portela ,2019) realizamos un método stepwise repetidas veces con diferentes archivos train sobre la muestra de mujeres y hombres. La salida incluye una tabla de frecuencias de los modelos que aparecen seleccionados en los diferentes archivos train. El criterio de selección puede ser AIC, SBC o BIC. Y además se puede poner un rango de semillas para que pruebe con varias a la vez cada vez que selecciona el archivo train. Hemos hecho una partición de los datos para la *macro randomselect* del 80% para un rango de semilla de 12345-12445 (100 semillas).

Con la macro *cruzada* (Portela, 2019*)* para regresiones normales con la variable dependiente de tipo intervalo, realizamos una validación cruzada repetida con varias semillas con los modelos de selección de variables obtenidos en SAS y R para ver cuál de ellos funciona mejor. La representación del error cuadrático medio de cada modelo se ha hecho en gráficos de caja.

A continuación, vamos a comentar los resultados obtenidos con las tablas y gráficos.

La selección de variables definitiva no se ha hecho de forma individual (ordenación o "importance") si no que se ha hecho por conjuntos, debido a la interdependencia de las variables en el modelo predictivo de los salarios. Los criterios iniciales que se han aplicado son Stepwise (AIC, BIC, SBC), Backward (AIC, BIC), Lasso.

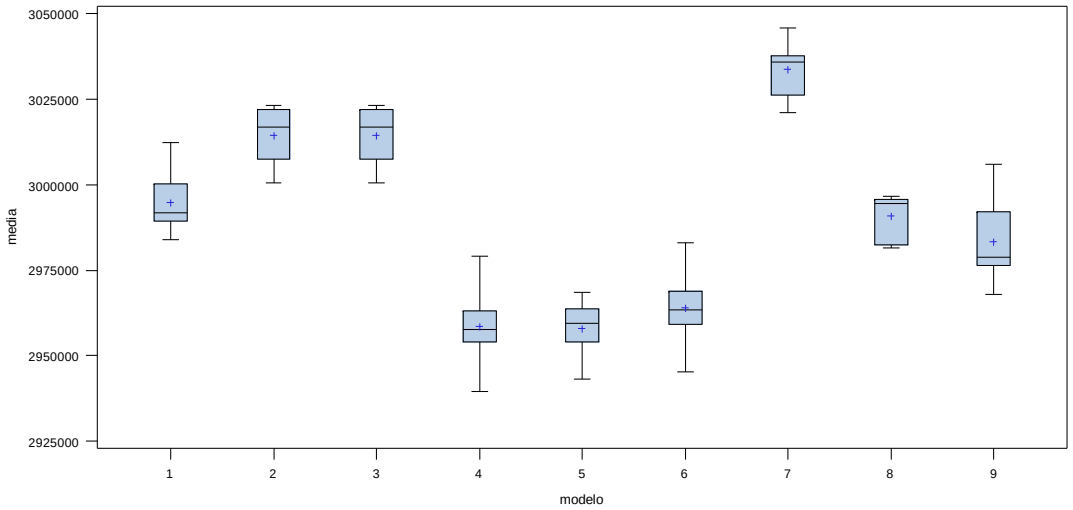
En SAS Base con la *macro randomselect* hemos obtenido 5 subconjuntos de datos para diferentes criterios de estadístico de selección de variables para las mujeres (modelos del 2 al 5). Y con R hemos obtenido también otros 4 subconjuntos de datos (modelos del 6 al 10).

Tabla 9 Selección de Variables para muestra de mujeres

Modelo	Criterio de selección	Conjunto de variables
Modelo 1	-	Todas las variables
Modelo 2	SBC	X9 X10 X12 X13 X26 X40 X41 X44 X53
Modelo 3	SBC	X10 X12 X13 X26 X40 X41 X43 X44 X53
Modelo 4	AIC	X4 X5 X8 X9 X10 X12 X13 X14 X20 X22 X26 X29 X34 X35 X40 X41 X42 X44 X48 X51 X53 X54 X55
Modelo 5	BIC	X4 X5 X8 X9 X10 X12 X13 X14 X20 X26 X28 X29 X34 X35 X40 X41 X42 X44 X48 X51 X53 X54
Modelo 6	Backward AIC	X4 X8 X9 X10 X12 X13 X14 X20 X22 X26 X28 X29 X35 X40 X41 X42 X44 X48 X49 X51 X53 X54
Modelo 7	Stepward BIC	X9 X10 X13 X35 X40 X41 X44 X51 X53
Modelo 8	Backward BIC	X4 X9 X10 X13 X20 X26 X35 X40 X41 X42 X44 X53
Modelo 9	Stepward AIC	X9 X10 X12 X13 X14 X20 X22 X25 X26 X35 X40 X41 X44 X45 X47 X48 X49 X51 X53 X54
Modelo 10	Lasso	*

* Las variables del modelo Lasso son difíciles de interpretar y además asciende a un total de 67 parámetros lo que lo no hace nada interesante este modelo para estudiar.

Ilustración 13. Comparación de los modelos de selección de variables obtenidas por SAS y R, muestra mujeres.



Con validación cruzada repetida se puede ver que los dos mejores modelos son el 4 y el 5 para las mujeres. El modelo 4 tiene menor error promedio pero mucha variabilidad, varianza, comparado con el modelo 5 que tiene un error promedio mínimamente superior pero sesgo y varianza más equilibrados. Además, el modelo 5 requiere una variable menos, lo que lo hace también más sencillo. Por lo que decidimos quedarnos con el modelo 5. El modelo elegido cuenta con las siguientes variables: continuas (X4 X8 X40 X41 X42) y categóricas (X5 X9 X10 X12 X13 X14 X20 X26 X28 X29 X34 X35 X44 X48 X51 X53 X54).

Destacamos que ninguno de los modelos obtenidos en el proceso de selección de variables para el conjunto de mujeres ha considerado la variable número de hijos, al contrario de lo que pasa con los hombres. podríamos crear un modelo que forzase a introducir la variable número de hijos, X55, pero no es el caso. Este dato aporta un matiz interesante al estudio.

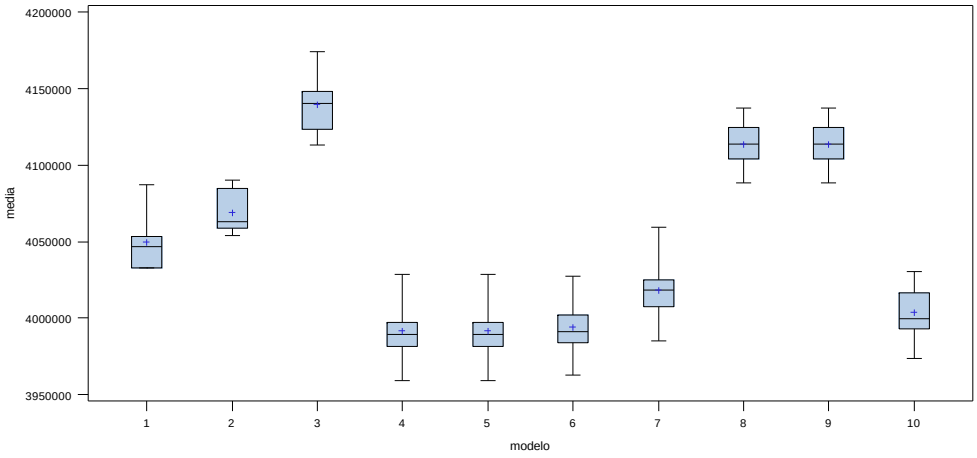
Tabla 10 Selección de Variables para muestra de hombres

Modelo	Criterio de selección	Conjunto de variables
Modelo 1	-	Todas las variables
Modelo 2	SBC	X4 X5 X9 X11 X26 X28 X40 X41 X42 X48 X53 X55
Modelo 3	SBC	X4 X5 X26 X28 X40 X41 X42 X43 X45 X48 X53 X55
Modelo 4	AIC	X4 X5 X9 X10 X11 X12 X17 X19 X20 X26 X28 X29 X30 X40 X41 X42 X44 X45 X47 X48 X53 X54 X55
Modelo 5	AIC	X4 X5 X10 X11 X12 X17 X19 X20 X26 X28 X29 X30 X40 X41 X42 X43 X44 X45 X47 X48 X53 X54 X55
Modelo 6	BIC	X4 X5 X9 X10 X11 X12 X19 X20 X26 X28 X29 X40 X41 X42 X44 X45 X47 X48 X53 X54 X55
Modelo 7	Stepward AIC	X4 X9 X10 X11 X16 X17 X20 X26 X28 X29 X30 X40 X41 X42 X45 X47 X48 X51 X53 X54 X55
Modelo 8	Stepward BIC	X4 X9 X26 X28 X40 X41 X42 X48 X51 X53
Modelo 9	Backward BIC	X4 X9 X26 X28 X40 X41 X42 X48 X51 X53
Modelo 10	Backward AIC	X4 X9 X12 X13 X14 X15 X16 X17 X18 X19 X20 X26 X28 X29 X30 X40 X41 X42 X44 X45 X47 X48 X51 X53 X54 X55
Modelo 11	Lasso	*

* Para simplificar la tabla no se han insertado las variables seleccionadas por estos modelos ya que no son interesantes para el estudio de validación cruzada para la comparación de los modelos.

Para los hombres se ha considerado los modelos que se ven en la Tabla 13. Los modelos del 2 al 6 son obtenidos en SAS y del 7 al 11 en R. En la Ilustración 14, con validación cruzada repetida, se puede observar que el mejor modelo es el 6, obtenido por SAS, caracterizado por tener mayor sesgo, menor varianza (a mayor varianza mayor sobreajuste del modelo) y menor número de variable respecto a los modelos 4 y 5, aunque tiene un error medio ligeramente superior, compensado por las demás características. El modelo elegido cuenta con las siguientes variables: continuas (X4 X19 X40 X41 X42 X45 X47) y categóricas (X5 X9 X10 X11 X12 X20 X26 X28 X29 X44 X48 X53 X54 X55).

Ilustración 14. Comparación de los modelos de selección de variables obtenidas por SAS y R, muestra hombres.



Una vez que ya tenemos definidos los sets de variables para hombres y mujeres damos paso a la modelización para obtener las ecuaciones de salarios.

10. Modelización

Una vez que ya hemos definido con que conjuntos de variables queremos trabajar, damos paso al punto más importante de este trabajo: la modelización. En este punto lo que pretendemos es encontrar y evaluar el mejor modelo con las técnicas de minería de datos aprendidas a lo largo del máster para nuestro conjunto de datos. Este procedimiento se va a aplicar sobre ambas muestras, mujeres y hombres.

En primer lugar, vamos a probar con diferentes modelos de regresión lineal, redes neuronales, árboles, bagging, random forest y gradient boosting y elegir el mejor de cada categoría. En segundo lugar, vamos a ver cuál es el mejor modelo de entre todas las categorías y por último aplicar la metodología de Oaxaca-Blinder y comprobar que el mejor de los modelos es el que consigue estimar mejor las predicciones de salarios así como el índice de discriminación.

10.1 Regresión Lineal.

Para ambas muestras y con R hemos aplicado el modelo de regresión que tiene como objetivo predecir el salario neto a partir de las m variables x_i independientes de los modelos seleccionados con anterioridad a través de la siguiente ecuación:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \epsilon \,, \qquad (12)$$

Donde y es el $salario\ neto$ que es una variable aleatoria continua, β_0 es el valor que toma la variable dependiente cuando todas las variables independientes valen 0, los β_m representan cuanto aumenta o disminuye la variable dependiente por cada incremento unitario de las variables independiente y ϵ es el error cometido o, equivalentemente, la parte del $salario\ neto$ que no queda explicada por las variables independientes.

Tabla 11. Resultados Regresión Lineal para mujeres y hombres.

	RMSE	Rsquared	RMSESD	RsquaredSD
Mujeres	1714.778	0.9854017	169.3218	0.002900925
Hombres	1994.162	0.9854372	123.4486	0.001717381

El modelo de regresión lineal funciona mejor para la muestra de hombres que para el de mujeres, pues tiene mayor r-cuadrado y menor desviación típica del error medio, aunque mayor error medio.

10.2 Redes Neuronales.

Este algoritmo es potente y robusto, pero necesita parametrización. Para las redes neuronales hemos usado SAS Base® y R®. En las redes uno de los parámetros más importante a controlar es el número de nodos ocultos. ¿Cuántos nodos ocultos necesitaríamos para cada uno de los data sets? Vamos a trabajar con 22 variables para la ecuación de mujeres y 21 variables para la ecuación de hombres. Dado el número de variables con el que contamos tendremos que adaptar el número de nodos basándonos

en este criterio: queremos como mínimo entre 20 y 30 observaciones por parámetro de la red. Para ello con las observaciones del conjunto de datos train (70%) calcularemos cuántos nodos es óptimo para una buena red neuronal. Para ello usaremos la siguiente fórmula:

$$Nº\ parámetros = h\ (k + 1) + h + 1\ , \qquad (13)$$

donde *h* son los nodos ocultos y *k* los nodos input (parámetros a introducir), tenemos que:

- Para el conjunto de datos de mujeres tenemos un total de 3452 observaciones para datos train. El número de parámetros, fijando un mínimo de 30 observaciones por parámetros, sería $\frac{3452}{30} = 115$ parámetros y con *k=22* tenemos*:* $115 = h(22 + 1) + h + 1 \rightarrow 115 - 1 = 24h \rightarrow h = 4\ o\ 5\ nodos$ *ocultos* podría estar bien*.* Si fijásemos como mínimo 20 observaciones por parámetro podríamos admitir hasta 7 u 8 nodos.
- Para el conjunto de datos de hombres tenemos un total de 3758 observaciones para datos train. El número de parámetros, fijando un mínimo de 30 observaciones por parámetros, sería $\frac{3758}{30} = 125$ parámetros y con *k=21* tenemos*:* $125 = h(21 + 1) + h + 1 \rightarrow 125 - 1 = 23h \rightarrow h = 5\ o\ 6\ nodos$ *ocultos* podría estar bien*.* Si fijásemos como mínimo 20 observaciones por parámetro podríamos admitir hasta 8 o 9 nodos.

Hemos probado varios modelos de redes neuronales explorando el early stopping, variando las funciones de activación, los algoritmos y el número de nodos ocultos. Para explorar el sesgo[20]y la varianza[21, 22] de los modelos hemos hecho validación cruzada repetida con varias semillas. Este método se ha aplicado a través de la *macro %cruzadas (Portela, 2019)* y la representación en boxplot.

Para ambos conjuntos de datos se ha hecho un estudio previo del early stopping variando el método y la función de activación de una red básica con el mínimo de nodos recomendados por la fórmula anterior usando dos macros en SAS. Una de las macros entrena la red para diferentes métodos (`LEVMAR QUANEW CONGRA BPROP DBLDOG TRUREG`) y la otra para distintas funciones de activación (`TANH LOG ARC LIN SIN SOF GAU`). Tras esta primera comprobación se confirma que el método preferido es Levmar con función de activación tangente hiperbólica para mujeres y arc para hombres, probando con varias semillas y porcentaje de muestra de datos train. Con estos dos parámetros las redes tienden a cero rápidamente y además los datos train y test quedan superpuestos lo que confirma la no necesidad del uso de early stopping. No obstante, hemos creado unas redes neuronales con valores de early stopping para comparar por validación cruzada con otras redes, pero el nivel de varianza mejoraba a consta de un

[20] Es el error promedio cometido por desfase de los valores verdaderos.

[21] Es la variabilidad de las predicciones del modelo al variar los datos utilizados para construirlo.

[22] En las redes el aumento del número de nodos aumenta el sesgo y reduce la varianza.

deterioro importante del sesgo, por lo que finalmente decidimos no establecer criterio de parada para ambos conjuntos de datos.

Con R y la librería *caret* hemos creado una función denominada *avnnet* que estima diferentes redes neuronales para cada conjunto de datos y nos indica cual sería la mejor red de acuerdo con el criterio de mayor r-cuadrado y menor RMSE.

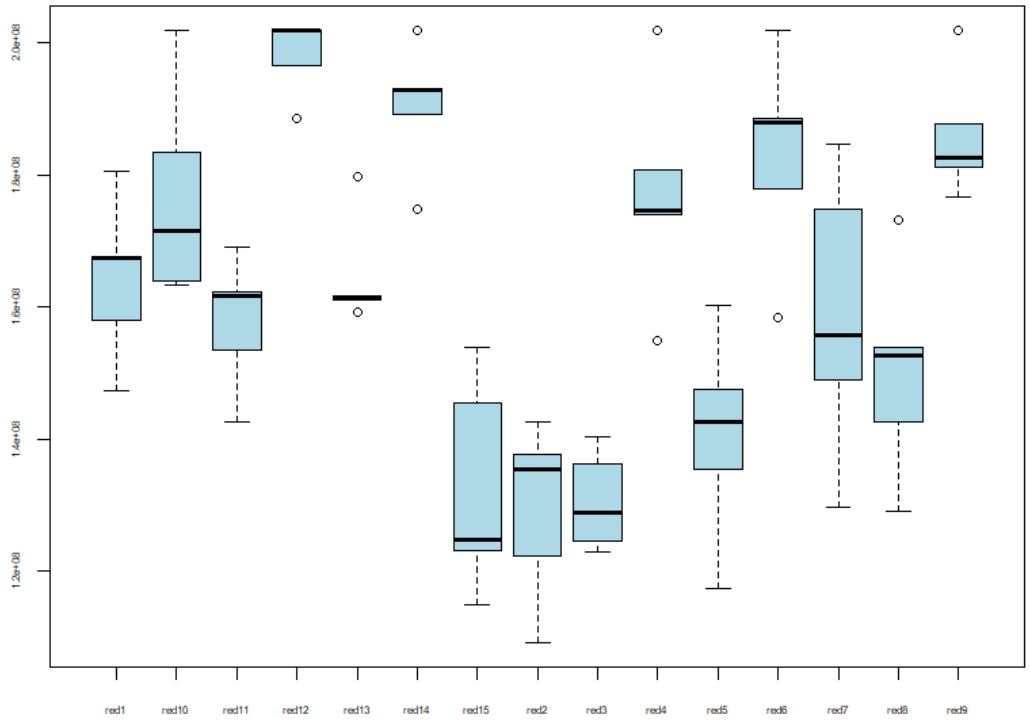
Con esta función de la librería *Caret* podemos jugar con el número de nodos y el decay.

Las diferentes redes probadas para la muestra de mujeres, según el software usado, se encuentran recogidas en la Tabla 12 y Tabla 13.

Tabla 12 Redes neuronales en R para muestra de mujeres.

Modelo	Nodos ocultos	Decay	Modelo	Nodos ocultos	Decay
Red 1	3	0.01	**Red 9**	6	0.01
Red 2	3	0.1	**Red 10**	6	0.1
Red 3	3	0.2	**Red 11**	6	0.2
Red 4	4	0.01	**Red 12**	7	0.01
Red 5	4	0.1	**Red 13**	7	0.2
Red 6	5	0.01	**Red 14**	8	0.01
Red 7	5	0.1	**Red 15**	8	0.2
Red 8	5	0.2			

Ilustración 15. Comparación por Validación Cruzada de las redes en R, muestra mujeres.

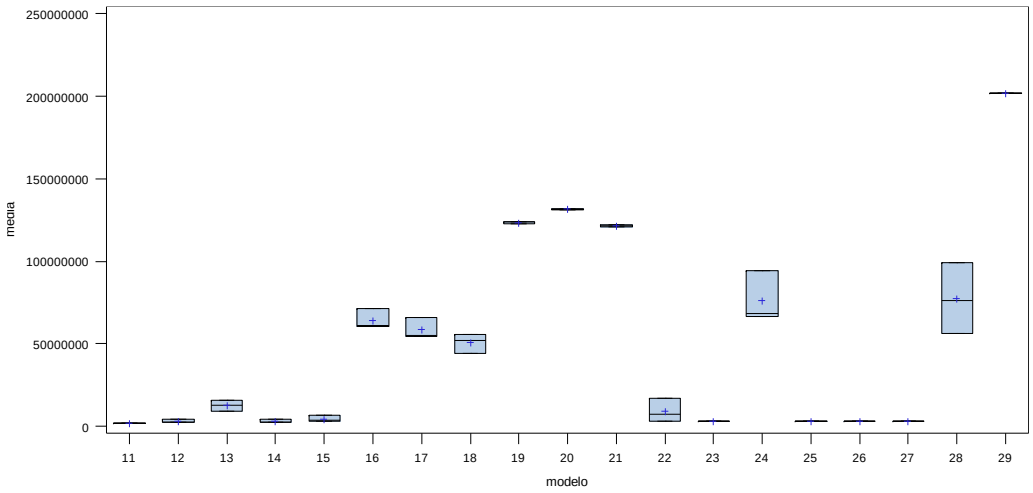


En R la mejor red es la 15, dado que tiene menor error RMSE en media y sesgo y varianza equilibrados, de los modelos comparados por validación cruzada repetida con la función *Cruzadaavnnet* y representados en la Ilustración 15. Esta red tiene 8 nodos y función de activación tanh por defecto y decay de 0.2, pero no algoritmo de optimización.

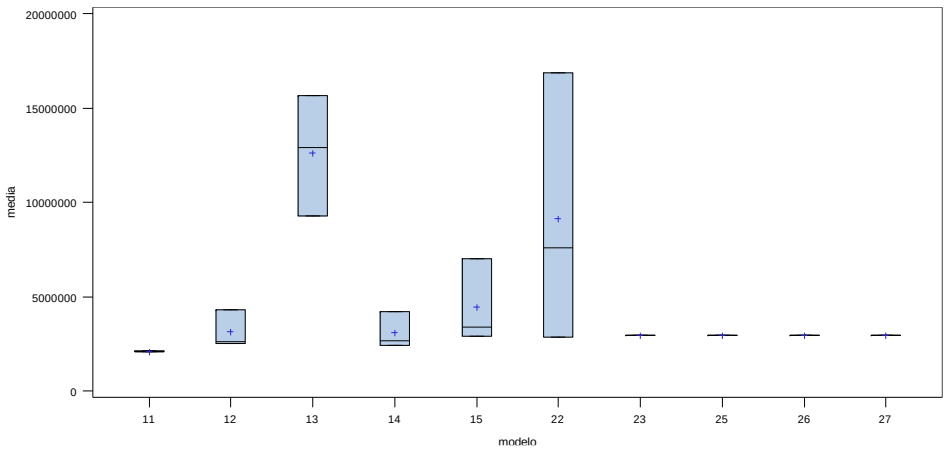
Tabla 13 Redes neuronales en SAS para muestra de mujeres.

Modelo	Nodos ocultos	Algoritmo	Función de activación	Early stopping
Modelo 11	4	Levmar	Tanh	
Modelo 12	5	Levmar	Tanh	
Modelo 13	6	Levmar	Tanh	
Modelo 14	7	Levmar	Tanh	
Modelo 15	8	Levmar	Tanh	
Modelo 16	5	Levmar	Tanh	40
Modelo 17	5	Levmar	Tanh	50
Modelo 18	5	Levmar	Tanh	60
Modelo 19	5	Bprop mom=0.2 learn=0.1	Tanh	
Modelo 20	6	Bprop mom=0.2 learn=0.1	Tanh	
Modelo 21	5	Bprop mom=0.8 learn=0.2	Tanh	
Modelo 22	5	Levmar	Log	
Modelo 23	3	Levmar	Lin	
Modelo 24	4	Quanew	Tanh	
Modelo 25	4	Levmar	Lin	
Modelo 26	5	Levmar	Lin	
Modelo 27	6	Levmar	Lin	
Modelo 28	5	Levmar	Gau	
Modelo 29	5	Bprop mom=0.2 learn=0.1	Gau	

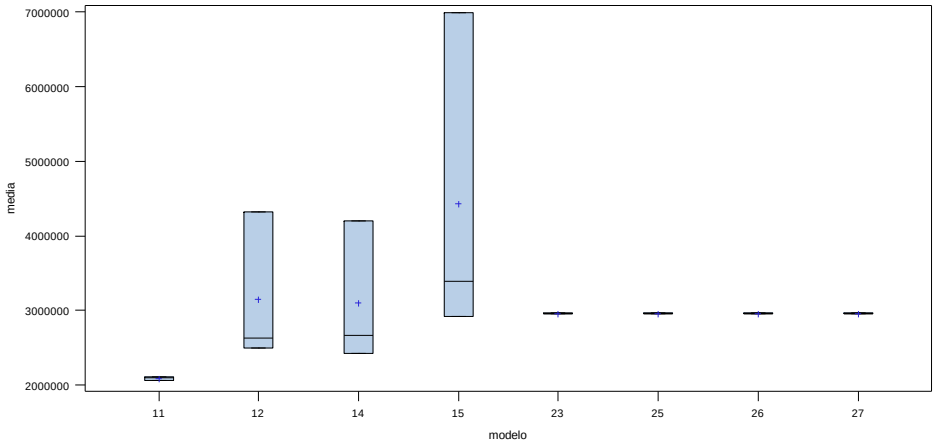
Ilustración 16. Comparación por Validación Cruzada de las Redes Neuronales en SAS para muestra de mujeres.



En el primer diagrama de cajas no se puede ver con claridad los modelos, por ello descartamos a los modelos del 16 al 21, el 24, 28 y 29 por tener un error medio superior al resto. Pasando a tener el siguiente diagrama que si nos permite evaluar y seleccionar la mejor red:



En este segundo gráfico, de nuevo no se ve con claridad, por lo quitamos a los modelos 13 y 22.



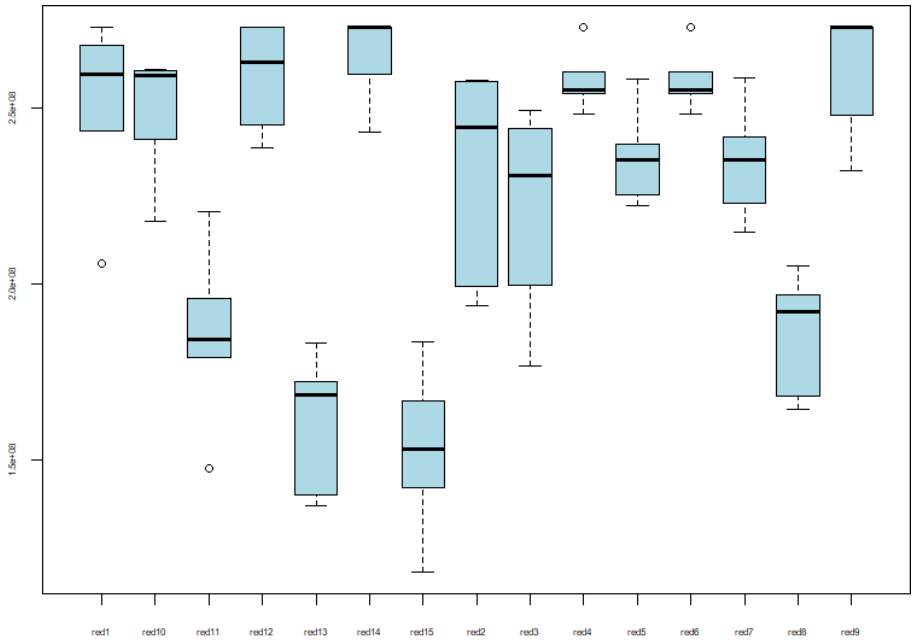
En este diagrama de cajas podemos ver que el mejor modelo sería la red 11 que comete un menor error medio. Se trata de una red de 4 nodos con algoritmo levmar y función de activación tangente hiperbólica.

Este mismo análisis descrito para la muestra de mujeres se ha repetido para la muestra de hombres. Las diferentes redes probadas se encuentran recogidas en la Tabla 14 y La mejor red neuronal obtenida con R para la muestra de hombres es la red 15 como se ve en el diagrama de cajas de la Ilustración 17. Esta red se caracteriza por estar formada por 10 nodos y un decay de 0.2. Además de tener sesgo y varianza proporcionales. Tabla 15.

Tabla 14 Redes Neuronales en R para muestra de hombres.

Modelo	Nodos ocultos	Decay
Red 1	5	0.01
Red 2	5	0.1
Red 3	5	0.2
Red 4	6	0.01
Red 5	6	0.1
Red 6	6	0.2
Red 7	7	0.1
Red 8	7	0.2
Red 9	8	0.01
Red 10	8	0.1
Red 11	8	0.2
Red 12	9	0.01
Red 13	9	0.2
Red 14	10	0.01
Red 15	10	0.2

Ilustración 17. Comparación por Validación Cruzada de las Redes Neuronales en R para muestra de hombres.

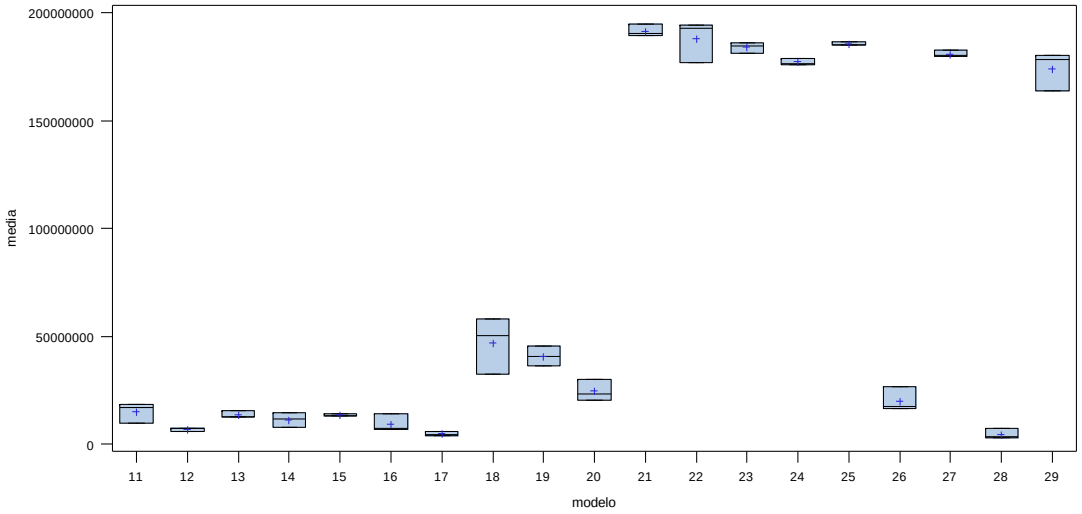


La mejor red neuronal obtenida con R para la muestra de hombres es la red 15 como se ve en el diagrama de cajas de la Ilustración 17. Esta red se caracteriza por estar formada por 10 nodos y un decay de 0.2. Además de tener sesgo y varianza proporcionales.

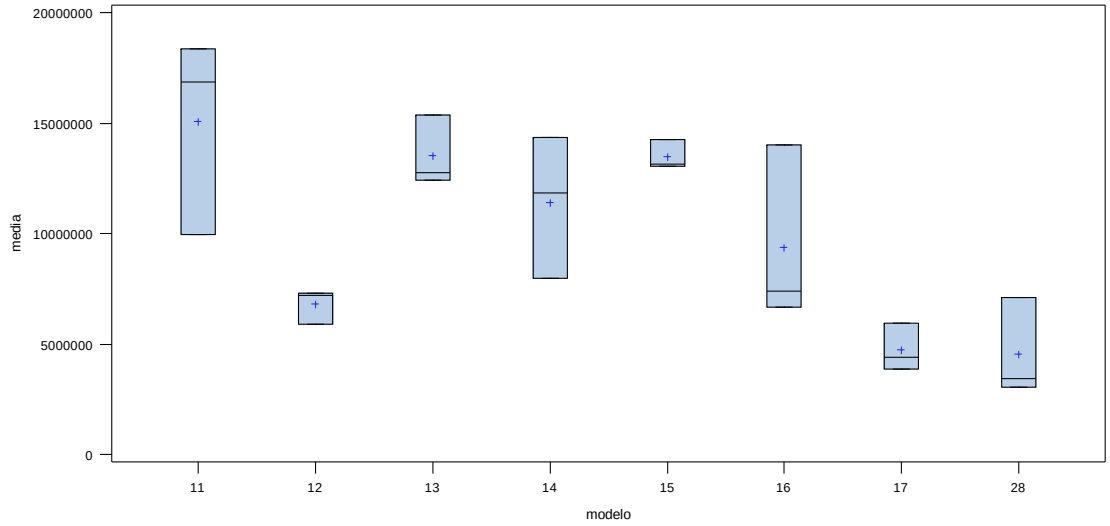
Tabla 15 Redes Neuronales probadas en SAS para muestra de hombres.

Modelo	Nodos ocultos	Algoritmo	Función de activación	Early stopping
Modelo 11	5	Levmar	Tanh	
Modelo 12	6	Levmar	Tanh	
Modelo 13	7	Levmar	Tanh	
Modelo 14	8	Levmar	Tanh	
Modelo 15	9	Levmar	Tanh	
Modelo 16	10	Levmar	Tanh	
Modelo 17	11	Levmar	Tanh	
Modelo 18	6	Levmar	Tanh	40
Modelo 19	8	Levmar	Tanh	50
Modelo 20	9	Levmar	Tanh	60
Modelo 21	5	Bprop mom=0.2 learn=0.1	Tanh	
Modelo 22	6	Bprop mom=0.2 learn=0.1	Log	
Modelo 23	7	Bprop mom=0.8 learn=0.2	Log	
Modelo 24	9	Bprop mom=0.2 learn=0.1	Tanh	
Modelo 25	9	Bprop mom=0.8 learn=0.2	Log	
Modelo 26	10	Quanew	Tanh	
Modelo 27	7	Bprop mom=0.8 learn=0.2	Arc	
Modelo 28	6	Levmar	Arc	
Modelo 29	10	Bprop mom=0.2 learn=0.1	Log	

Ilustración 18. Validación Cruzada para Redes Neuronales en SAS para muestra de hombres.



Para SAS ocurre lo mismo que en el caso de las mujeres. Los modelos del 18 al 27 y el 29 no nos permiten visualizar correctamente a los demás modelos por tener un error medio superior al de las demás redes. Por ello los suprimimos para visualizar mejor.



De este modo se puede ver que el modelo ganador sería la red 28, una red de 6 nodos, con algoritmo Levmar y función de activación arc.

Hemos visto que las redes en principio tienen la varianza muy acotada, pero el sesgo es muy diferente según la red seleccionada. Por este motivo hemos elegido los modelos con más nodos, por que de este modo se mejora en sesgo.

10.1 Árboles de regresión.

Los árboles de regresión los hemos hecho en R con *cruzada arbol continua* (Portela, 2019). Hemos probado el modelo más sencillo de árbol binario. A los cuales les hemos pedido que tengan un número mínimo de obervaciones por nodo de 20. Los resultados obtenidos se pueden ver en las tabla a continuación.

Tabla 16. Resultados Árbol de Regresión para mujeres y hombres.

	RMSE	Rsquared	MAE
Mujeres	4638.265	0.8914871	2714.563
Hombres	4709.323	0.9183746	2775.099

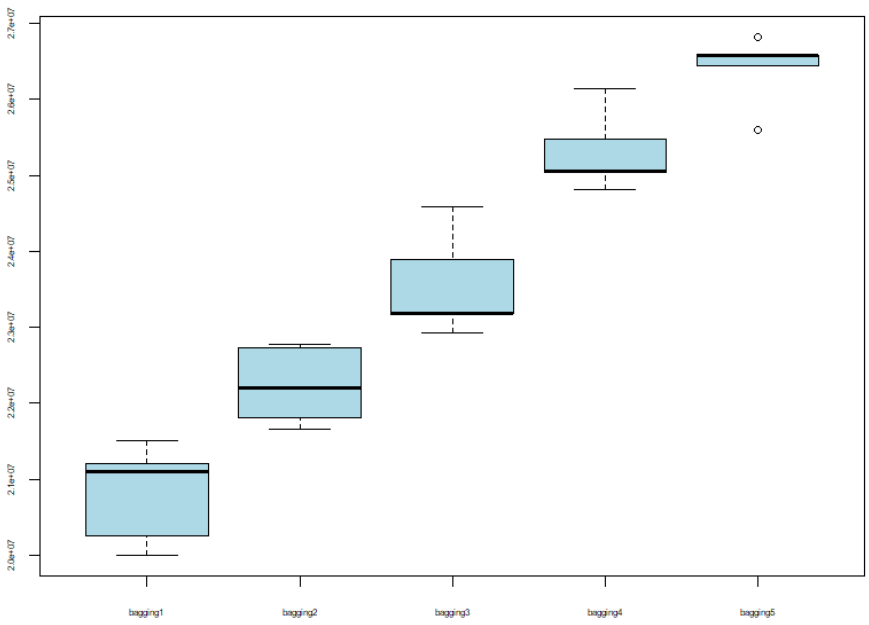
El árbol para la muestra de hombres se ajusta mucho mejor que para las mujeres. Tiene un r-cuadrado superior a consta de un mayor error cuadrático medio.

10.2 Bagging.

Para los modelos de árboles de tipo bagging hemos utilizado SAS con la macro *%cruzadarandomforest* (Portela, 2019) y R con el programa *cruzada rf continua* (Portela, 2019). Se han probado una gran variedad de modelos. En SAS con la macro hemos podido modificar diferentes parámetros tales como: el porcentaje de muestra para la construcción de árboles, el tamaño mínimo de observaciones para las hojas finales, la profundidad máxima del árbol y el p-valor para creación de nuevos nodos. Mientras que para el programa de R hemos tomado una muestra del 70% y un tamaño mínimo de observaciones para las hojas finales.

Tabla 17. Modelos Bagging entrenados en R para muestra mujeres.

Modelo	Nodesize	Ntree
Bagging 1	10	200
Bagging 2	15	200
Bagging 3	20	200
Bagging 4	25	300
Bagging 5	30	300

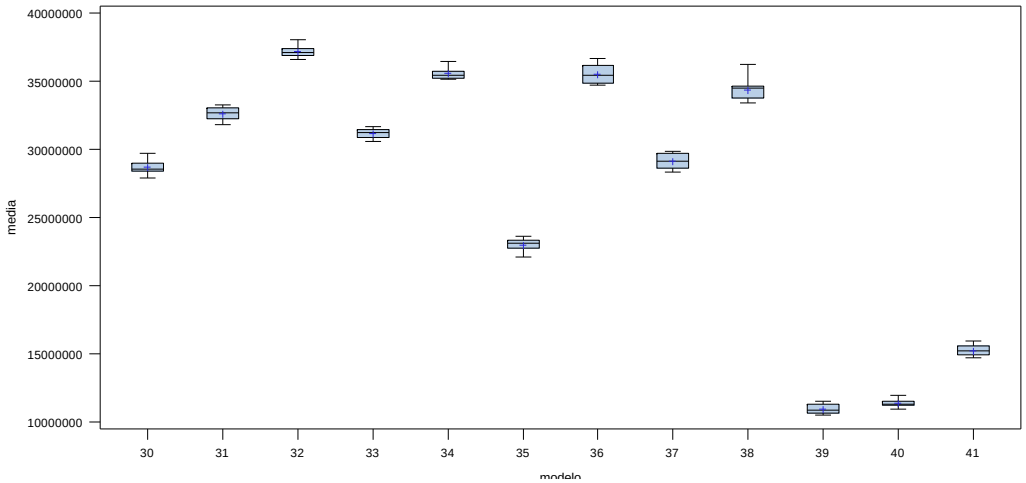


En R el mejor modelo para las mujeres es bagging 1, dado que tiene menor error medio y parte de un modelo más sencillo, sólo tiene 10 observaciones por nodo.

Tabla 18 Modelos de Bagging en SAS para muestra de mujeres.

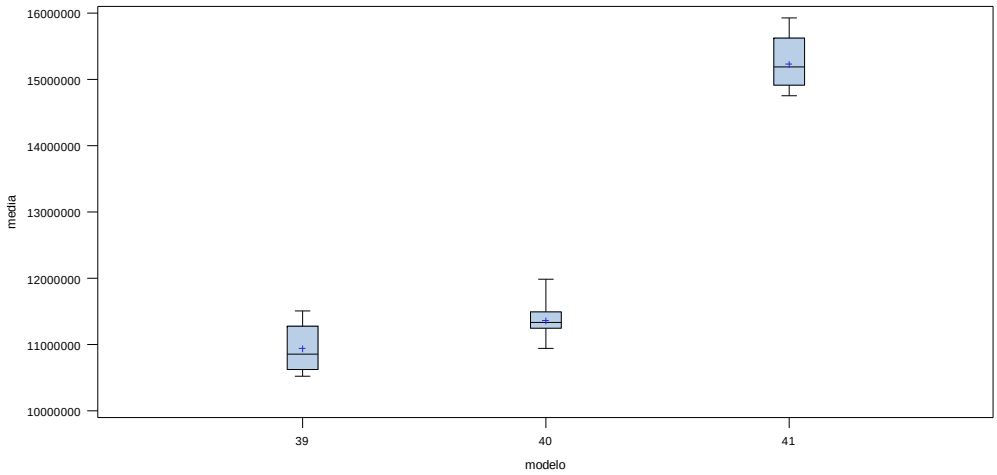
Modelo	Porcenbag	Tamhoja	Maxdepth	p-valor
Modelo 30	0.5	20	10	0.1
Modelo 31	0.5	25	10	0.1
Modelo 32	0.5	30	10	0.1
Modelo 33	0.7	20	6	0.1
Modelo 34	0.7	30	6	0.1
Modelo 35	0.8	20	10	0.05
Modelo 36	0.8	30	6	0.05
Modelo 37	0.8	30	10	0.15
Modelo 38	0.8	40	10	0.2
Modelo 39	0.8	5	10	0.2
Modelo 40	0.8	5	10	0.1
Modelo 41	0.8	10	10	0.1

Ilustración 19. Comparación por Validación Cruzada de los modelos probados de bagging para mujeres.



En la

Ilustración 19 vemos que los posibles mejores modelos son del 39 al 41, eliminamos los modelos del 30 al 38 para visualizar mejor.

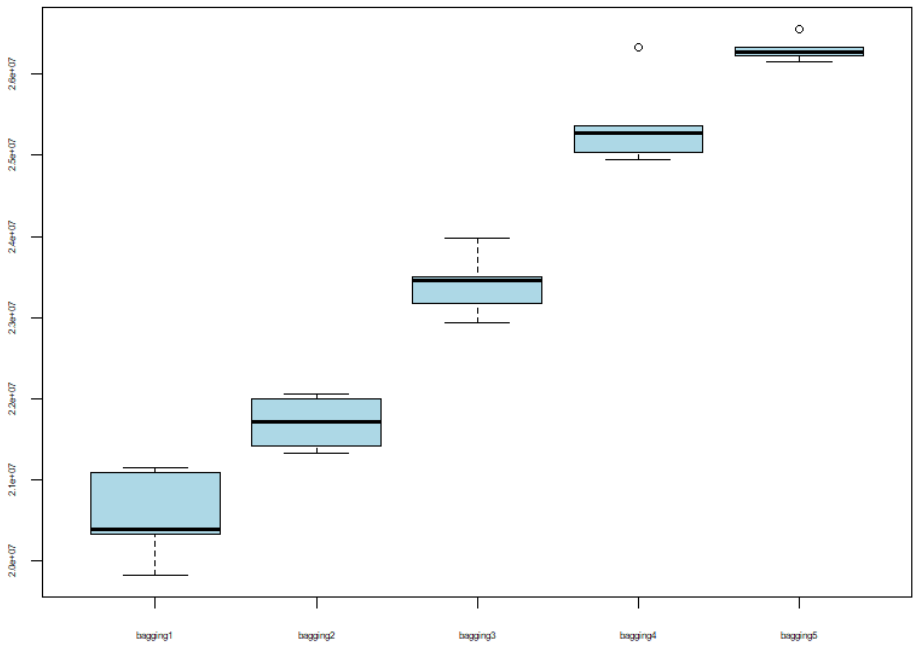


El mejor modelo de bagging para el conjunto de datos de mujeres en SAS es el modelo 39, por tener el menor error medio y además de tener el sesgo y varianza compensados.

Se trata de un modelo bagging que toma un 80% de muestra, 5 observaciones por hoja final, una profundidad máxima de 10 con un p-valor de 0.2, es un modelo muy agresivo.

Tabla 19. Modelos de bagging probados en R para muestra de hombres.

Modelo	Nodesize	Ntree
Bagging 1	10	200
Bagging 2	15	200
Bagging 3	20	200
Bagging 4	25	300
Bagging 5	30	300

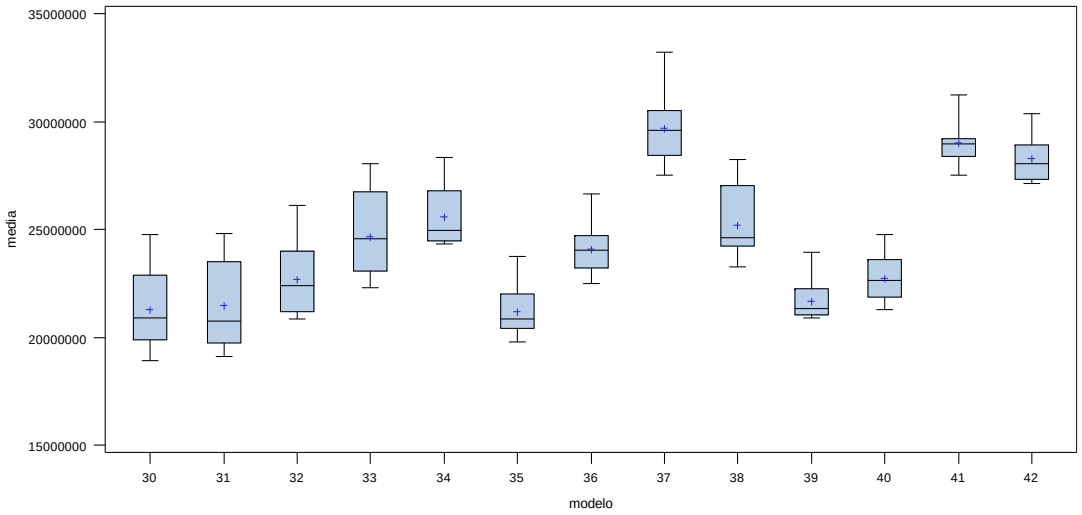


En R el mejor modelo para los hombres es bagging 1, dado que tiene menor error medio y parte de un modelo más sencillo, sólo tiene 10 observaciones por nodo.

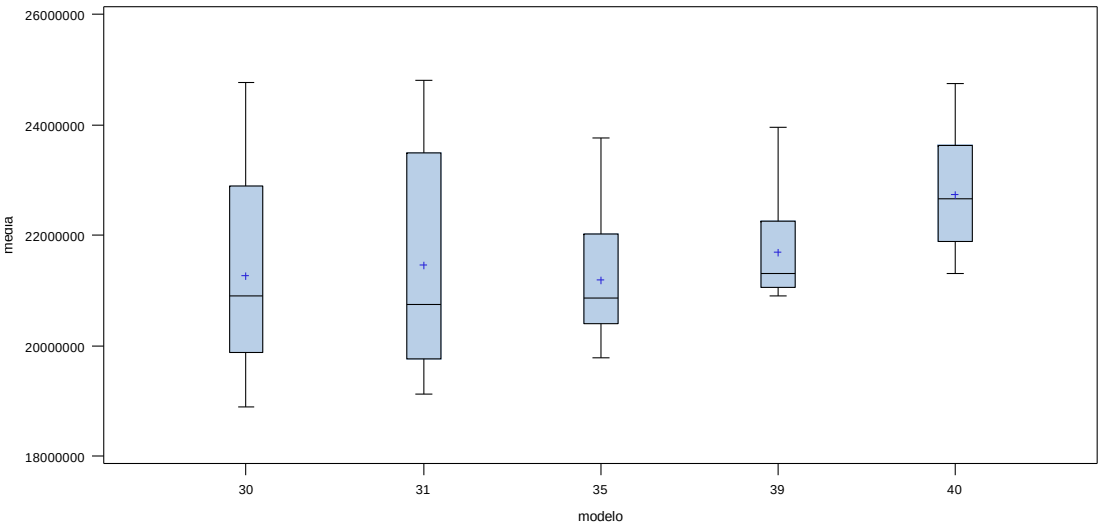
Tabla 20 Modelos de bagging en SAS para muestra de hombres.

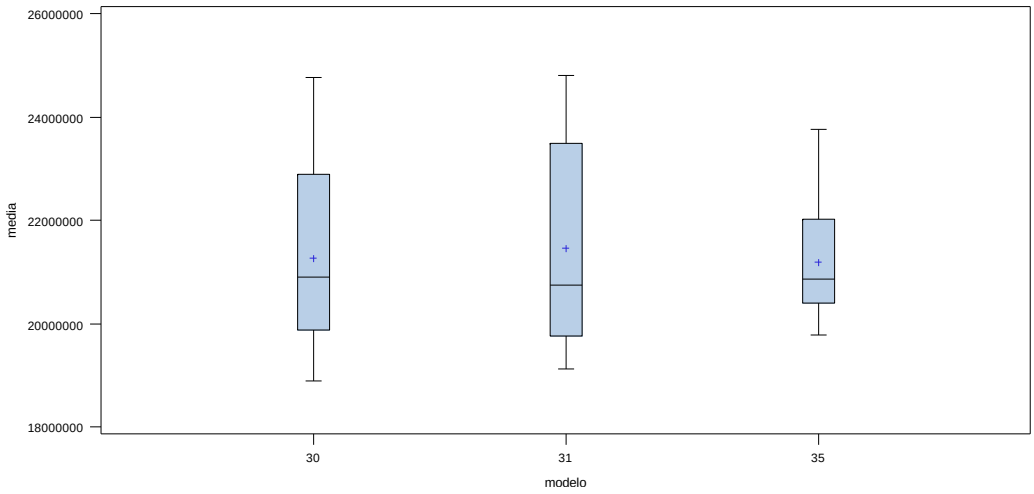
Modelo	Porcenbag	Tamhoja	Maxdepth	p-valor
Modelo 30	0.8	10	6	0.05
Modelo 31	0.8	10	6	0.1
Modelo 32	0.8	15	6	0.1
Modelo 33	0.8	20	6	0.1
Modelo 34	0.8	30	10	0.15
Modelo 35	0.7	20	10	0.2
Modelo 36	0.7	25	10	0.1
Modelo 37	0.7	30	6	0.05
Modelo 38	0.7	20	6	0.15
Modelo 39	0.5	15	10	0.1
Modelo 40	0.5	10	6	0.1
Modelo 41	0.5	25	10	0.2
Modelo 42	0.5	20	6	0.2

Ilustración 20 Comparación por Validación Cruzada de los modelos de Bagging en SAS para muestra de hombres.



En la Ilustración 20 vemos que los posibles mejores modelos de bagging son el 30, 31, 35, 39 y 40, los dibujamos aparte.





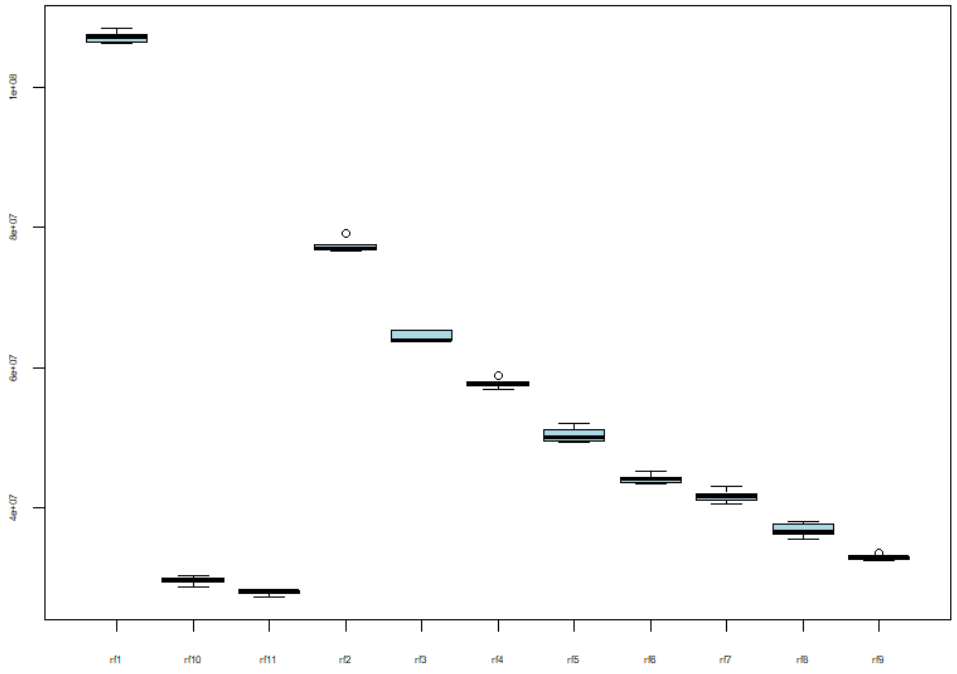
El mejor modelo de bagging para el conjunto de datos de hombres por menor error promedio es el modelo 31, que toma un 80% de muestra, 10 observaciones por hoja final, una profundidad máxima de 6 con un p-valor de 0.1.

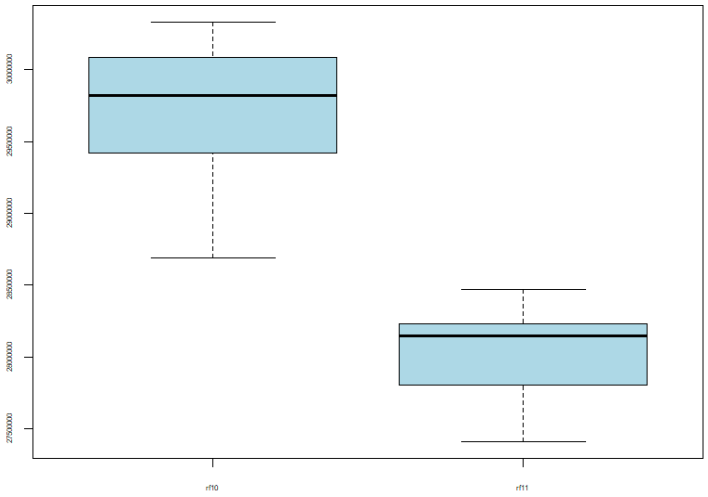
10.3 Random Forest.

Como sabemos Random Forest es una ampliación del modelo de bagging. Por lo que se usa la misma macro de SAS y el mismo programa de R, pero con el matiz de que variamos el parámetro referente al número de variables a sortear en el inicio de cada nodo. Además, en R hemos usado la función *rf*, que entrena un modelo para diferentes números de variables, probando desde el modelo más sencillo hasta el que contiene a todas las variables (bagging). Para el caso de las mujeres hemos obtenido por esta función que el mejor modelo es aquel que usa 23 variables (bagging), aun así hemos probado otros modelos intermedios y comparado por validación cruzada, modelos que se encuentran resumidos en la Tabla 21 y en la Tabla 22.

Tabla 21 Modelos de Random Forest probados en R para muestra de mujeres.

Modelo	Nodesize	Mtry
Rf1	35	2
Rf2	10	4
Rf3	10	6
Rf4	20	8
Rf5	20	10
Rf6	20	12
Rf7	30	14
Rf8	30	16
Rf9	30	18
Rf10	30	20
Rf 11	30	21



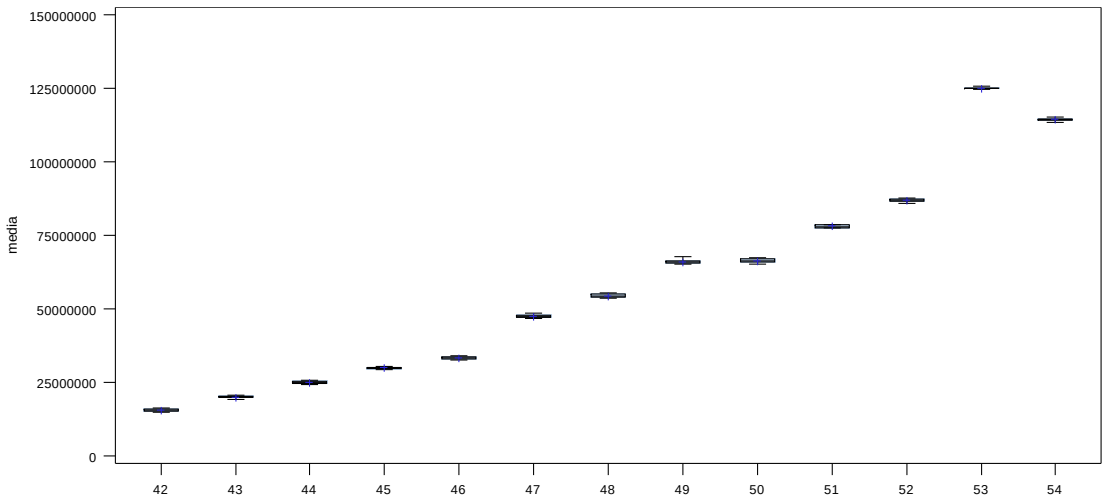


En R hemos obtenido que los dos mejores modelos para las mujeres son Rf10 y Rf11, ampliamos estos dos boxplots y vemos que en efecto el mejor modelo es Rf11. Un modelo con un tamaño de nodo de 30 observaciones y 21 variables. Lo que nos había sugerido la función *rf,* es preferible bagging a random forest.

Tabla 22 Modelos de Random Forest en SAS para muestra de mujeres.

Modelo	Variables	% bag	Tamhoja	Maxdepth	P valor
Modelo 42	21	0.8	10	10	0.1
Modelo 43	20	0.8	15	10	0.1
Modelo 44	19	0.8	20	10	0.1
Modelo 45	18	0.8	25	10	0.1
Modelo 46	16	0.8	25	10	0.1
Modelo 47	15	0.8	25	6	0.15
Modelo 48	13	0.7	25	6	0.15
Modelo 49	10	0.8	30	6	0.2
Modelo 50	8	0.8	35	10	0.15
Modelo 51	6	0.7	25	10	0.15
Modelo 52	4	0.8	25	10	0.2
Modelo 53	2	0.7	25	6	0.1
Modelo 54	2	0.8	10	10	0.1

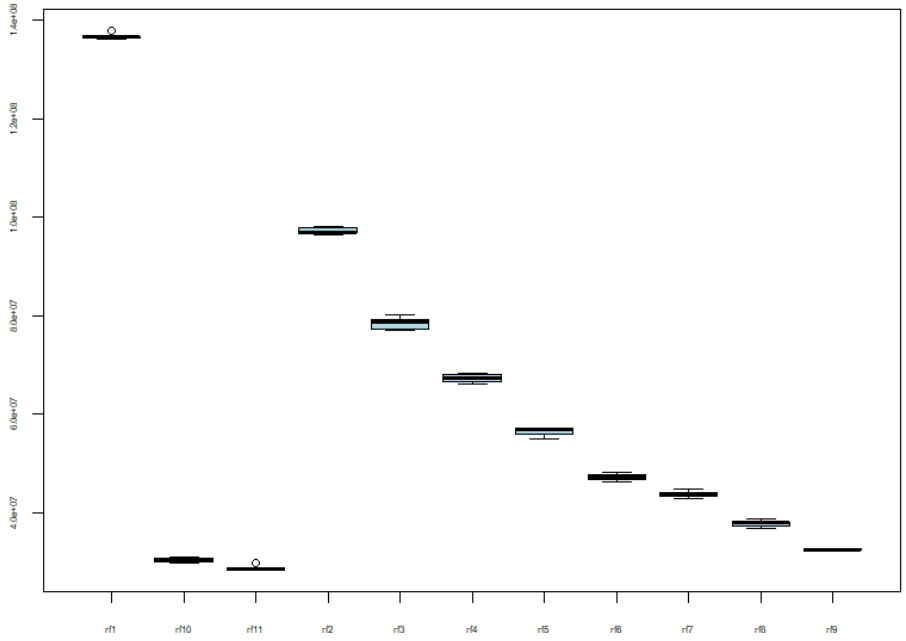
Ilustración 21 Comparación por Validación Cruzada de los modelos de Random Forest en SAS para muestra de mujeres.

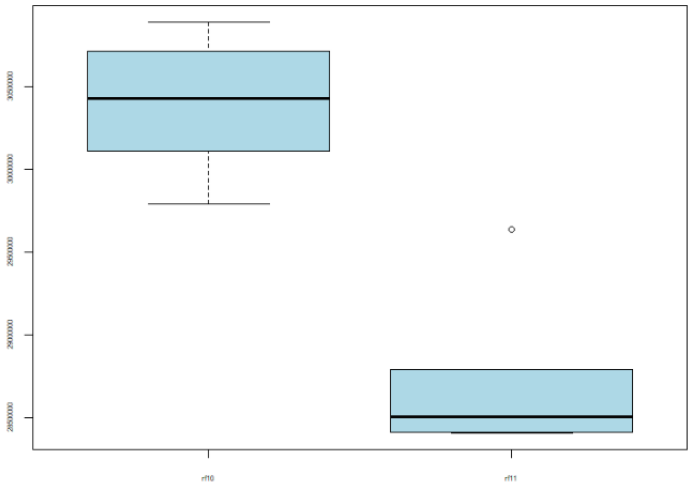


El mejor modelo de Random Forest para las mujeres es el 42 que se crea estableciendo que en el inicio de cada nodo se utilicen 21 variables diferentes, con un porcentaje de la muestra del 80%, un tamaño mínimo de observaciones en últimos nodos de 10 observaciones, una profundidad máxima de 10 y un p-valor de 0,1.

Tabla 23 Modelos de Random Forest en R para muestra de hombres.

Modelo	Nodesize	Mtry
Rf1	35	2
Rf2	10	4
Rf3	10	6
Rf4	20	8
Rf5	20	10
Rf6	20	12
Rf7	30	14
Rf8	30	16
Rf9	30	18
Rf10	30	19
Rf 11	30	20



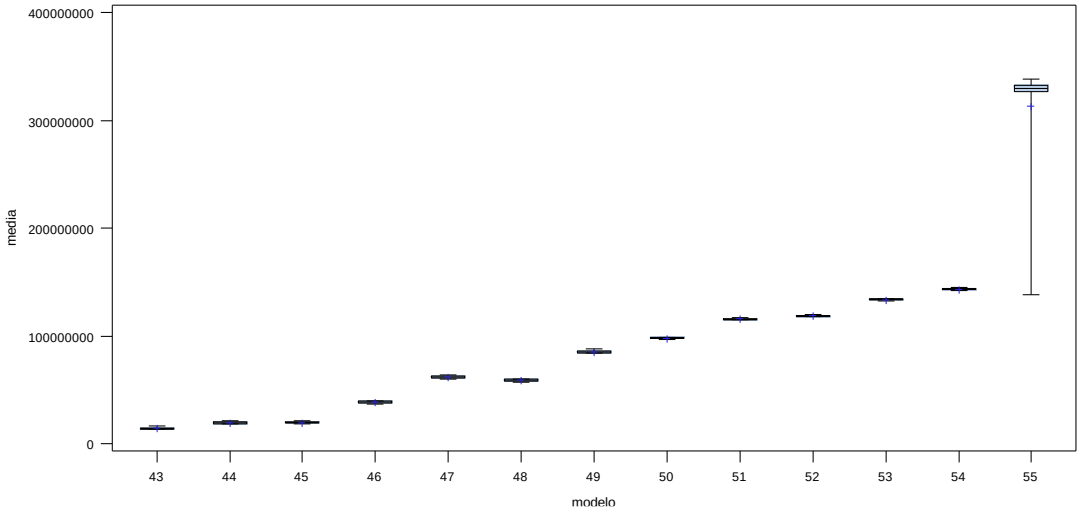


Para los hombres hemos obtenido que el mejor modelo de Random Forest en R es el Rf11 con un tamaño mínimos de observaciones en último nodo de 30 y 20 variables.

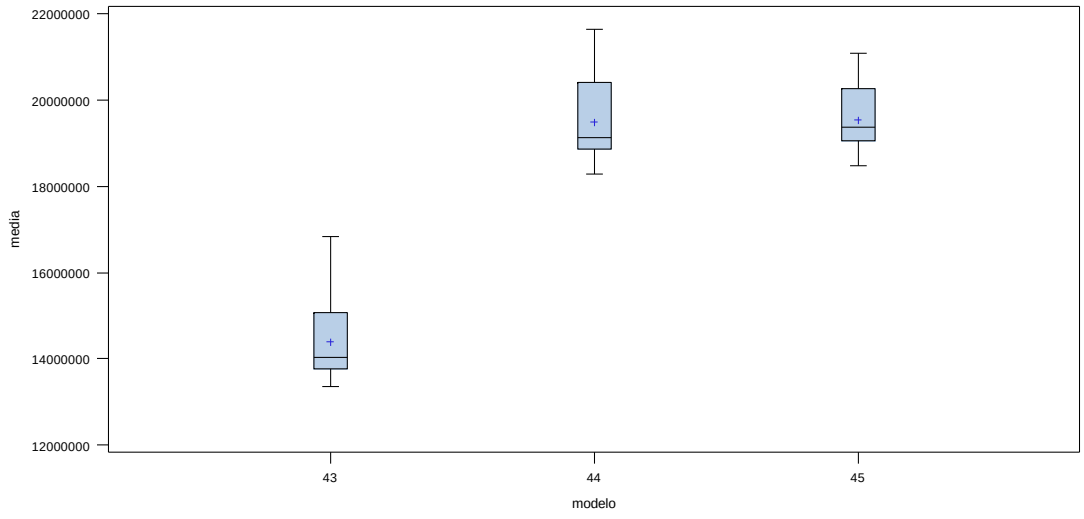
Tabla 24 Modelos de Random Forest en SAS para muestra de hombres.

Modelo	Variables	% bag	Tamhoja	Maxdepth	p-valor
Modelo 43	20	0.8	10	10	0.1
Modelo 44	18	0.8	15	10	0.1
Modelo 45	16	0.8	10	10	0.1
Modelo 46	14	0.8	25	10	0.1
Modelo 47	12	0.8	25	6	0.1
Modelo 48	10	0.8	20	10	0.15
Modelo 49	8	0.8	20	6	0.15
Modelo 50	6	0.8	30	10	0.1
Modelo 51	5	0.7	35	10	0.1
Modelo 52	4	0.7	25	10	0.15
Modelo 53	3	0.7	25	6	0.2
Modelo 54	2	0.8	25	10	0.2
Modelo 55	2	0.8	35	4	0.1

Ilustración 22 Comparación por Validación Cruzada de modelos de Random Forest en SAS para muestra de hombres.



Parece que los modelos 43, 44 y 45 son los que tienen menor error. Eliminamos los modelos del 46 al 55 para visualizar mejor.



El mejor modelo de Random Forest en SAS para los hombres es el 43 que se crea estableciendo que en el inicio de cada nodo se utilicen 20 variables diferentes, con un porcentaje de la muestra del 80%, un tamaño mínimo de observaciones en últimos nodos de 10 observaciones, una profundidad máxima de 10 y un p-valor de 0,1.

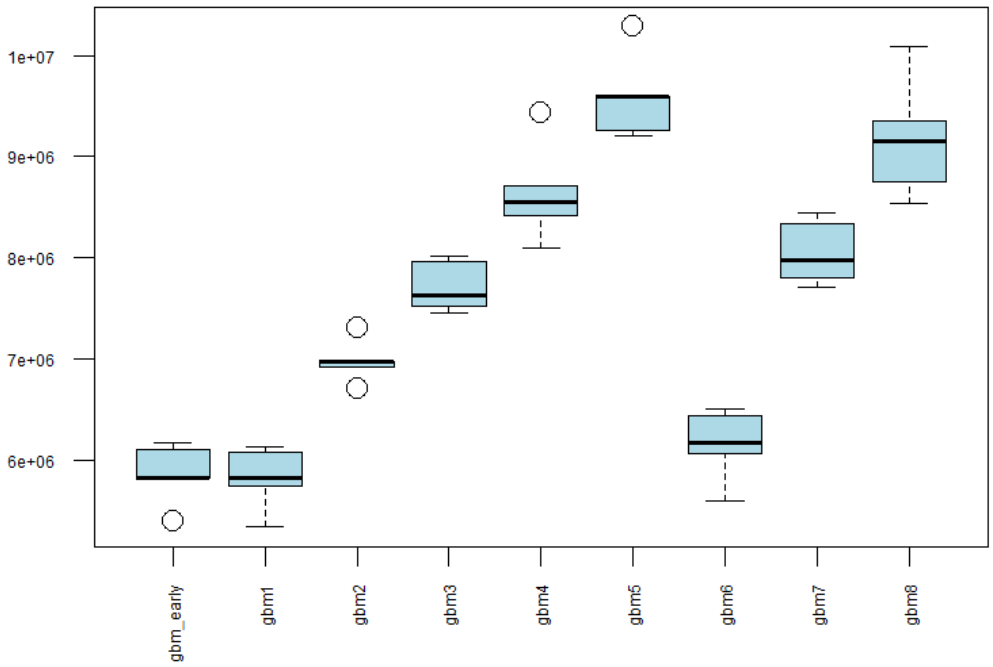
10.4 Gradient Boosting.

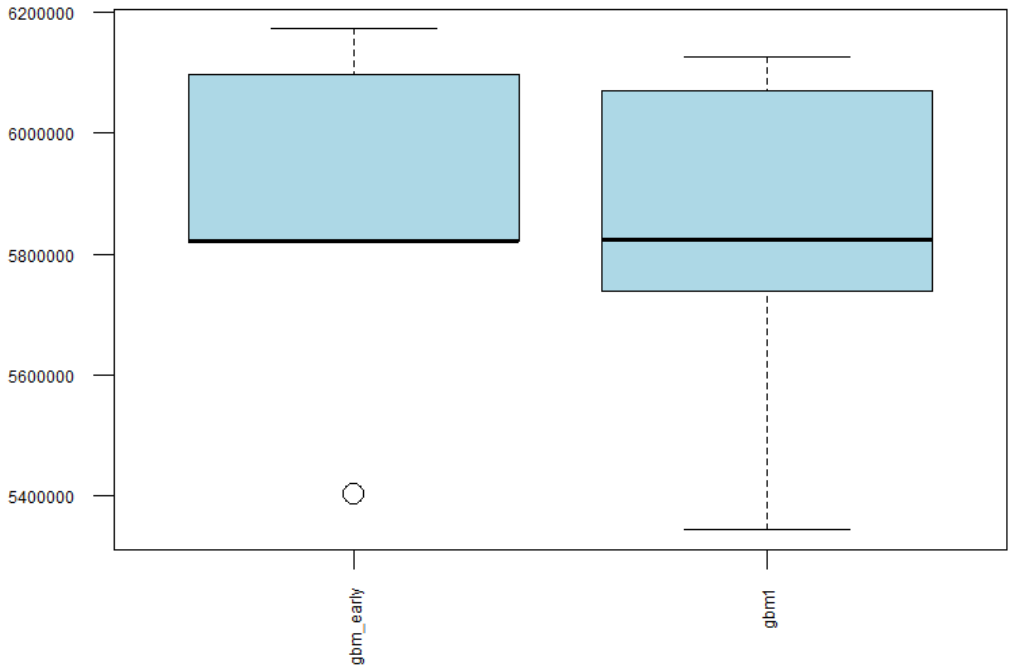
Este último algoritmo, también basado en árboles, introduce un parámetro llamado shrinkage que es una tasa de aprendizaje que gradúa cuánto se va corrigiendo el error en cada iteración. Para tasas de aprendizaje muy pequeñas son necesarias más iteraciones, mientras que para tasas moderadas el número de iteraciones necesarias es inferior. En R con el programa ***cruzada gbm continua*** hemos usado la función gbm generando una rejilla para que pruebe con distintos valores de los parámetros shrinkage y ntress (número de iteraciones) y una modificación para que haga early stopping para el número de iteraciones.

Tabla 25 Modelos Gradient Boosting en R para muestra de mujeres.

Modelo	nminobsinnode	Shrinkage	Ntrees
Gbm 1	15	0.05	5000
Gbm 2	20	0.05	5000
Gbm 3	25	0.05	5000
Gbm 4	30	0.05	5000
Gbm 5	35	0.05	5000
Gbm 6	15	0.1	5000
Gbm 7	25	0.1	5000
Gbm 8	30	0.1	5000
Gbm early	15	0.05	8000

Ilustración 23 Comparación por Validación Cruzada de modelos de Gradient Boosting en R para muestra de mujeres.



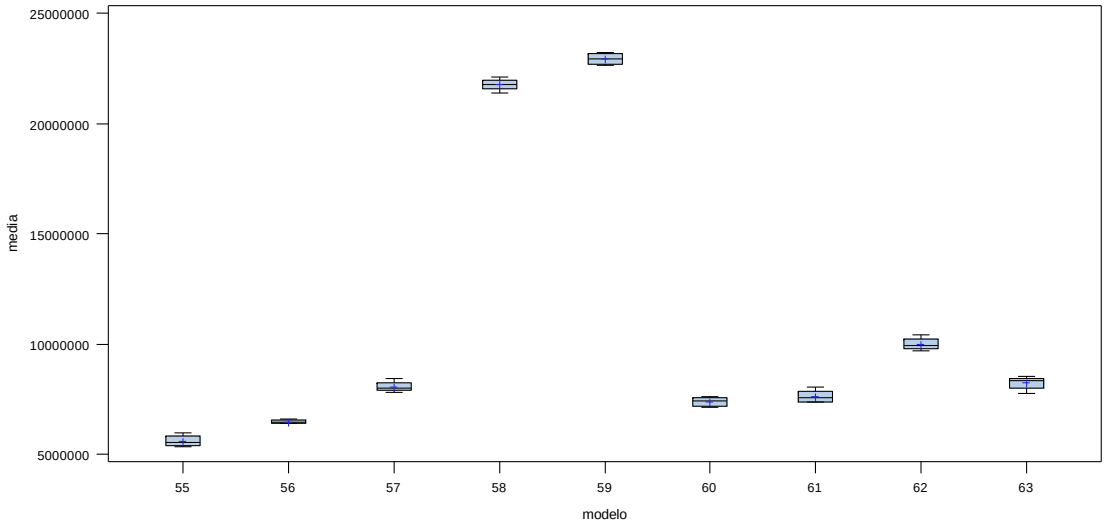


Los dos mejores modelos en R para mujeres son gbm1 y gbmearly. Comparándolos individualmente vemos que tienen el mismo error promedio, pero varían su sesgo y varianza, en este caso es mejor gbm1, dado que tiene más sesgo y menor varianza.

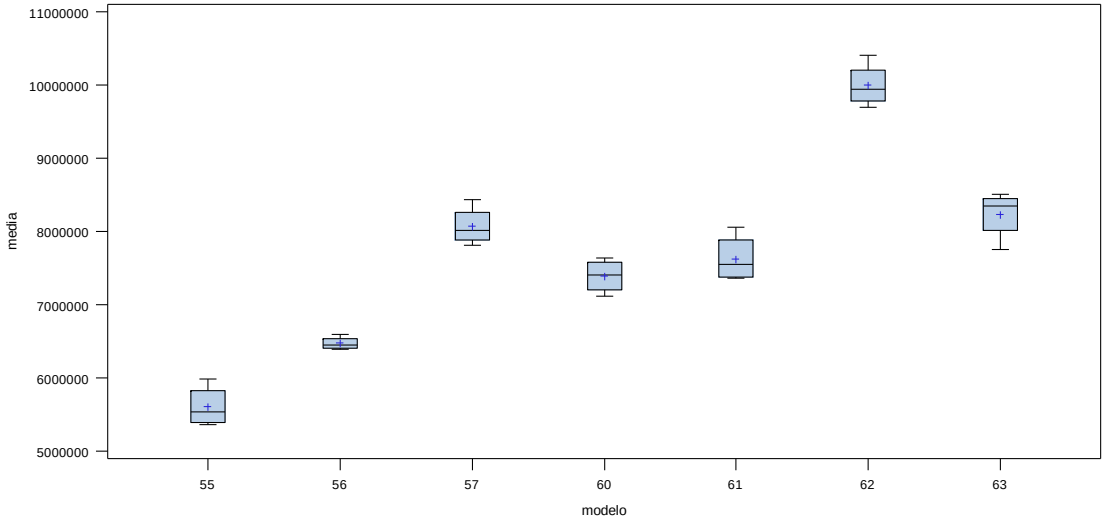
Tabla 26. Modelos de Gradient Boosting en SAS para muestra de mujeres.

Modelo	Leafsize	Shrink	Minobs	Mincatsize	Maxdepth
Modelo 55	10	0.05	10	10	4
Modelo 56	15	0.1	15	15	4
Modelo 57	20	0.2	30	10	4
Modelo 58	25	0.01	20	10	4
Modelo 59	30	0.01	30	20	4
Modelo 60	15	0.2	30	10	6
Modelo 61	20	0.1	30	-	6
Modelo 62	35	0.1	15	10	6
Modelo 63	25	0.1	25	-	4

Ilustración 24. Comparación por Validación Cruzada de modelos de Gradient Boosting en SAS para muestra de mujeres.



Parece que el modelo 55 va bien, pero vamos a eliminar a los modelos 58 y 59 para visualizar mejor los gráficos de caja, pues estos modelos tienen un elevado error medio e impiden evaluar correctamente a los demás.

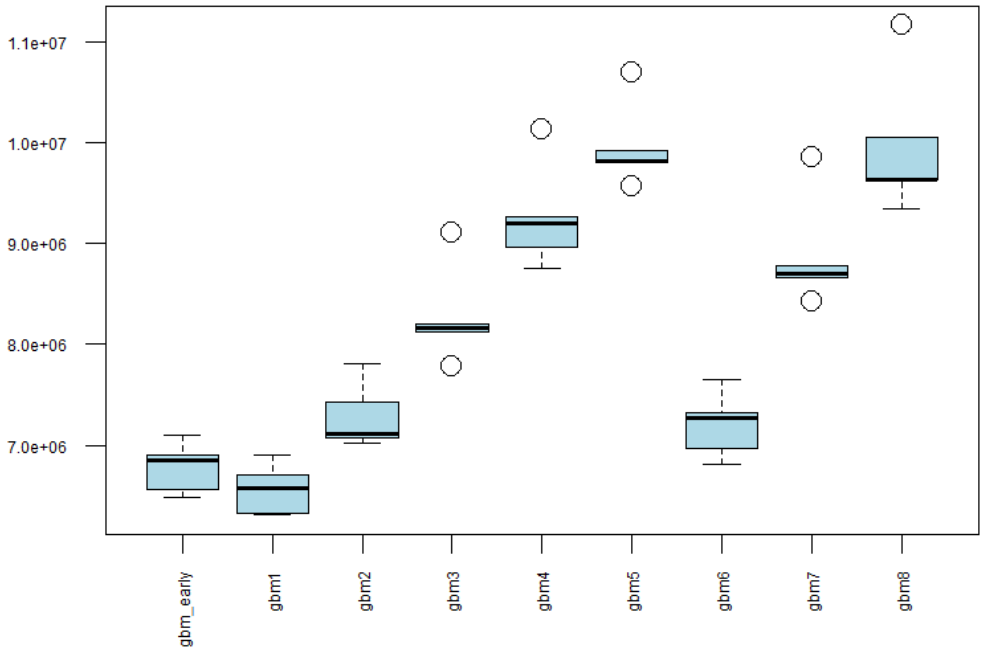


El mejor modelo de Gradient Boosting para las mujeres es el 55 creado con un tamaño de hoja final de 10, con 10 observaciones mínimo para el inicio de cada nodo, una tasa de aprendizaje del 0.05 y 10 observaciones como mínimo para variable categórica.

Tabla 27. Modelos de Gradient Boosting en R para muestra hombres.

Modelo	Nminobsinnode	Shrinkage	Ntrees
Gbm 1	15	0.03	5000
Gbm 2	20	0.05	5000
Gbm 3	25	0.05	5000
Gbm 4	30	0.05	5000
Gbm 5	35	0.05	5000
Gbm 6	15	0.1	5000
Gbm 7	25	0.1	5000
Gbm 8	30	0.1	5000
Gbm early	15	0.05	5000

Ilustración 25. Comparación por Validación Cruzada modelos Gradient Boosting en R para muestra de hombres.

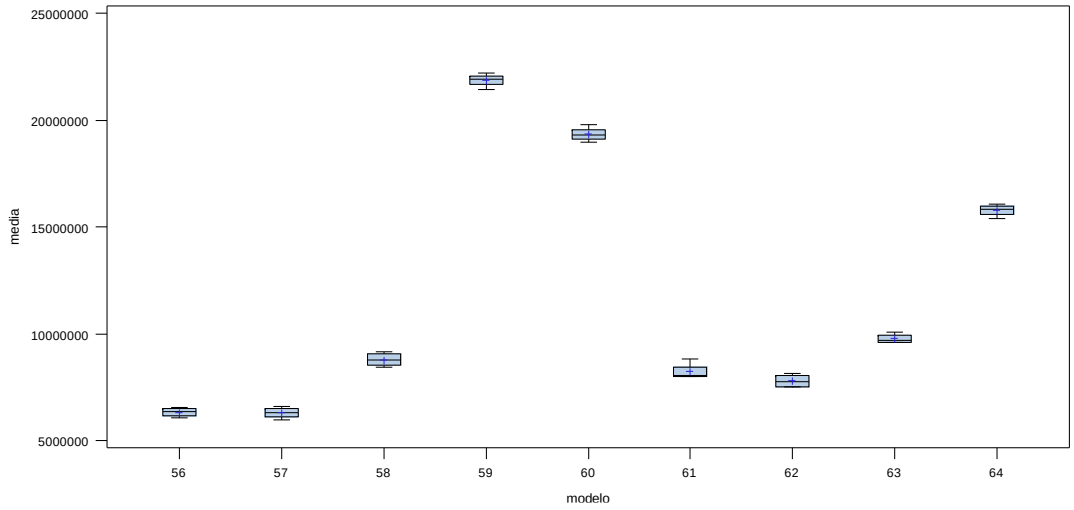


El mejor modelo en R para hombres es gbm1, es el modelo que tiene menor error y varianza.

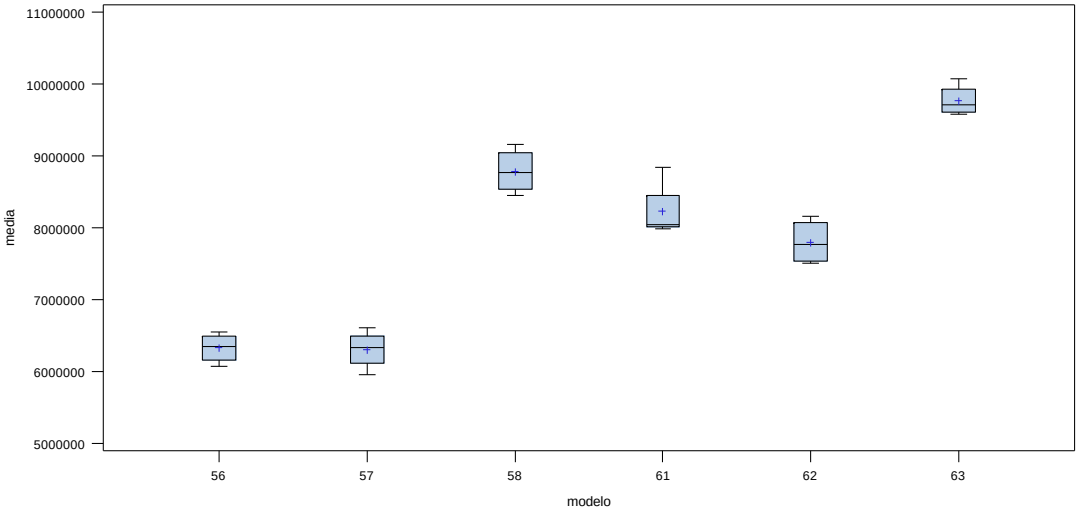
Tabla 28. Modelos de Gradient Boosting en SAS para muestra hombres.

Modelo	Leafsize	Shrink	Mincatsize	Minobs	Maxdepth
Modelo 56	10	0.05	10	10	4
Modelo 57	10	0.1	20	20	4
Modelo 58	20	0.2	30	30	4
Modelo 59	30	0.01	10	20	4
Modelo 60	20	0.01	10	30	4
Modelo 61	15	0.2	10	30	4
Modelo 62	20	0.1	10	30	6
Modelo 63	30	0.1	10	25	6
Modelo 64	30	0.01	10	10	6

Ilustración 26. Comparación por Validación Cruzada de modelos de Gradient Boosting en SAS para muestra de hombres.



Eliminamos los modelos 59, 60 y 64 para visualizar mejor los gráficos de caja, pues estos modelos tienen un elevado error medio e impiden evaluar correctamente a los demás.



El mejor modelo de Gradient Boosting para los hombres es el 57 creado con un tamaño de hoja final de 10, con 20 observaciones mínimo para el inicio de cada nodo, una tasa de aprendizaje del 0.1, 20 observaciones como mínimo para variable categórica, 300 iteraciones y una profundidad máxima de 4.

11. Selección del mejor modelo de Estimación para la Ecuación de Salarios.

Damos paso a la sección más importante de este trabajo: la Estimación de la mejor ecuación de salarios para los hombres y para las mujeres. Hemos probado 6 algoritmos de *Machine Learning* diferentes para intentar buscar aquel que resuma de forma más fiable la evolución de los salarios de acuerdo con las características personales y laborales de cada individuo con datos train. Ahora con los datos test vamos a comparar los modelos para cada segmento de la población y evaluar qué software, SAS o R, predice mejor teniendo en cuenta la diferente arquitectura de los algoritmos entre ellos y las diferentes opciones de parametrización.

Tabla 29. Resumen de modelos.

Modelo	Mujeres	Hombres
Regresión Lineal	Por defecto, en R	Por defecto, en R
Red Neuronal	SAS: 4 nodos, Levmar, Tanh	SAS: 6 nodos, Levmar, Arc
	R: 8 nodos, 0.2 decay	R: 10 nodos, 0.2 decay
Árbol de regresión	Por defecto, en R	Por defecto, en R
Bagging	SAS: 80% muestra, 5 observaciones, máxima profundidad 10, p-valor=0.2	SAS: 80% muestra, 10 observaciones, máxima profundidad 6, p-valor=0.1
	R: tamaño de nodo 10, 200 árboles	R: tamaño de nodo 15, 200 árboles
Random Forest	SAS: 21 variables, 80% muestra, 10 observaciones, máxima profundidad 10, p-valor=0.1	SAS: 20 variables, 80% muestra, 10 observaciones, máxima profundidad 10, p-valor=0.1
	R: tamaño de nodo 30, 21 variables	R: tamaño de nodo 30, 20 variables
Gradient Boosting	SAS: 10 tamaño de hoja final, 10 observaciones inicio nodo, 10 observaciones por categórica, 0.05 tasa aprendizaje, máxima profundidad 4	SAS: 10 tamaño de hoja final, 20 observaciones inicio nodo, 20 observaciones por categórica, 0.1 tasa aprendizaje, máxima profundidad 4
	R: 15 observaciones inicio nodo, 0.05 tasa de aprendizaje, 5000 iteraciones	R: 15 observaciones inicio nodo, 0.03 tasa de aprendizaje, 5000 iteraciones

Hemos comparado estos modelos entre sí por validación cruzada repetida y hemos obtenido que el mejor modelo de machine learning para cada una de nuestras poblaciones es Gradient Boosting en R y la red neuronal en SAS (modelo 65 en Ilustración 28). Por lo tanto, el algoritmo de gradient boosting y la red neuronal podrían hacerle competencia a la clásica regresión lineal empleada para la estimación de ecuaciones de salarios.

Ilustración 27. Comparación modelos Machine Learning en R.

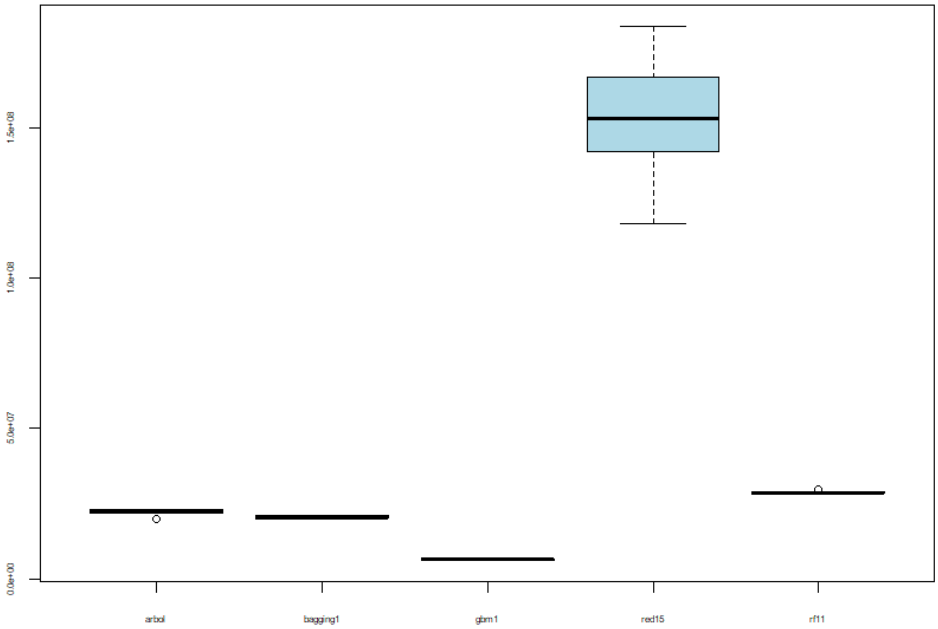
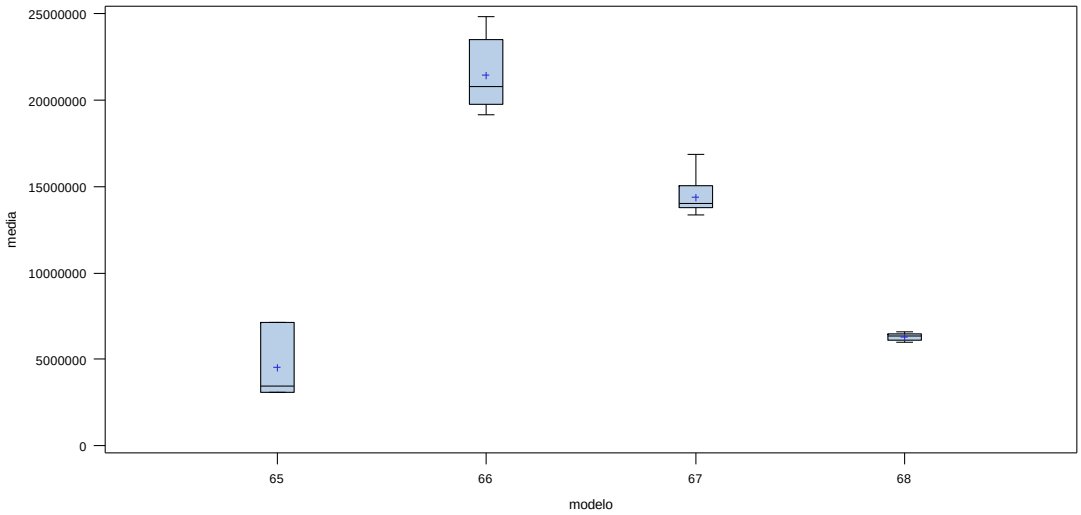


Ilustración 28. Comparación modelos Machine Learning en SAS.



NOTA: Donde el modelo 65 se corresponde con la red neuronal, el 66 con bagging, el 67 con random forest y el 68 con gradient boosting.

11.1 Aplicación de la metodología de Oaxaca-Blinder.

Recordemos que la metodología de Oaxaca-Blinder trata de estimar sendas ecuaciones de salarios para hombres y mujeres. Las sendas de ecuaciones las hemos obtenido para ambos segmentos de la población con diferentes algoritmos. Ahora lo que haremos será testear las ecuaciones de salarios con la partición de datos test que habíamos reservado al inicio del estudio y calcular las predicciones de los nuevos individuos para cada uno de los modelos. De tal modo que tendremos:

$$W_H = f^H(X_1 \dots X_K) \quad (14)$$

$$W_M = f^M(X_1 \dots X_K) \quad (15)$$

Donde W es el salario real neto mensual percibido por cada individuo *(M* para las mujeres y *H* para los hombres) para sus respectivas funciones en cada uno de los modelos.

Con las ecuaciones (*14*) y (*15*) podemos estimar el salario a percibir por cada individuo de acuerdo con sus características individuales en cada uno de los modelos estudiados. Pero lo verdaderamente interesante de este trabajo es estimar cómo sería retribuida una mujer de acuerdo con la ecuación de salarios masculina y viceversa. Para ello planteamos las siguientes ecuaciones:

$$W_{H^*} = f^M(X_1 \dots X_K) \quad (16)$$

$$W_{M^*} = f^H(X_1 \dots X_K) \quad (17)$$

Una vez que conozcamos las predicciones de los salarios a percibir por cada uno de los individuos de acuerdo con la ecuación de salarios de la población del sexo opuesto, manteniendo todo lo demás constante, podremos obtener el porcentaje de incremento salarial que debería recibir una mujer para igualarse al salario percibido si fuese un hombre:

$$\%\ incremento\ salarial\ a\ recibir\ las\ mujeres^{23} = \frac{W_{M^*} - W_M}{W_M} \quad (18)$$

Para ello hemos calculado las predicciones de la regresión lineal y de árbol de regresión en R junto con los mejores modelos obtenidos en R de redes neuronales, bagging, random forest y gradient boosting. En SAS sólo se han calculado las predicciones para los mejores modelos obtenidos en SAS para las redes neuronales, bagging, random forest y gradient boosting, dichas predicciones se han guardado en un data set con el que se ha trabajado posteriormente en R.

[23] Esta estimación se ha calculado única y exclusivamente en R Studio© a partir de las predicciones obtenidas en SAS© y dicho programa.

11.2 Resultados predictivos.

Antes de dar paso a las predicciones de los salarios de mujeres y hombres aplicando los datos test a los modelos de machine learning seleccionados hemos calculado cuál es el salario medio de los datos test para cada grupo de la población, como se ve en la

Tabla 30. Salario Neto Medio Anual para datos test.

Conjunto datos	Salario medio
Datos test mujeres	19607.50
Datos test hombres	23318.68

Ahora podemos comparar los salarios medios predichos por cada técnica con los que se tienen en los datos test.

Tabla 31. Predicción de Salarios Medios por modelo.

	Mujeres	Hombres
Modelo	**Predicción Salario Medio en R®**	**Predicción Salario Medio en R®**
Regresión lineal	19655.88	23320.53
Red Neuronal	18561.08	23760.31
Árbol	19446.74	23358.29
Bagging	19586.11	23330.57
Random Forest	19524.78	23345.83
Gradient Boosting	19633.89	23304.12

Comparando los resultados obtenidos en la Tabla 31 con los de la Tabla 30, podemos decir que los mejores modelos en R serían la regresión lineal, bagging y random forest por la mayor proximidad al dato medio.

En la Tabla 32 hemos representado el incremento salarial medio en porcentaje que debería de recibir una mujer para conseguir igualar su salario al salario que percibiría si sus características fuesen evaluadas según la ecuación de salarios masculina. Se ha obtenido la predicción para cada individuo de la población estudiado, pero dado que no se puede representar a todos, son miles, hemos optado por poner el valor promedio del resultado obtenido para la fórmula (18).

Tabla 32. Incremento salarial necesario para igualdad de salarios.

Modelo	Incremento salarial (%)
Regresión Lineal	4.200463%
Red Neuronal	4.093696%
Árbol	9.901924%
Bagging	4.205272%
Random Forest	4.635568%
Gradient Boosting	138.3058%

Los resultados predictivos muestran de nuevo que los mejores modelos que se ajustan son regresión lineal, red neuronal, bagging y random forest. Los dispares resultados de las predicciones se deben a las diferentes configuraciones de cada uno de los algoritmos,

lo que nos puede estar indicando que debemos mejorar aún más cada uno de los métodos, bien sea con las diferentes características a variar, introducir o extraer variables, etc.

En SAS se han calculado también las predicciones, así como el incremento salarial, pero se obtenían datos muy dispares, incrementos entorno al 138%, es decir, resultados parecidos al de gradient boosting en R. Por ello se ha decidido no introducirlos en este trabajo.

Los siguientes gráficos son de predicción de los modelos en SAS, dónde parecía que los modelos se ajustaban bastante bien a los datos test, pero en el análisis posterior dado a la dificultad de tratamiento de las salidas han impendido comprobar fehacientemente el incremento salarial a percibir. Pero muestran que son modelos buenos para nuestros datos estudiados.

Ilustración 29. Predicción Salario Mujeres evaluadas como Hombres. Redes Neuronales.

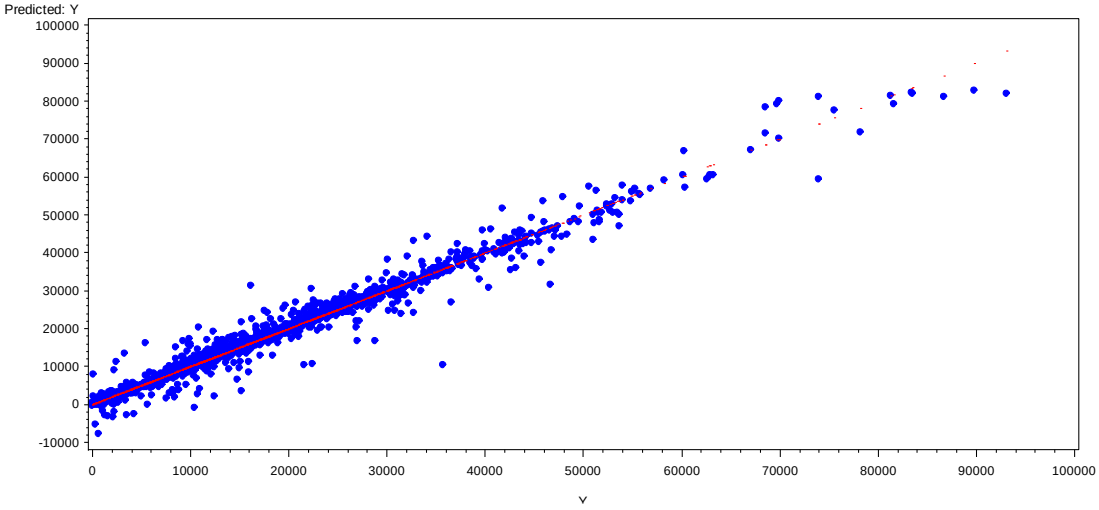


Ilustración 30. Predicción Salario Mujeres evaluadas como Hombres. Gradient Boosting

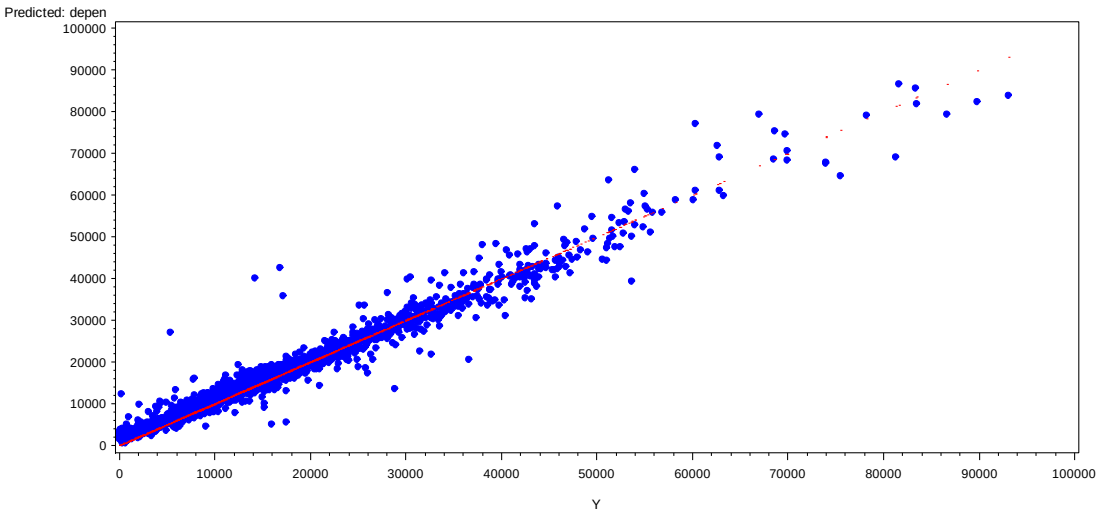


Ilustración 31. Predicción Salario Mujeres evaluadas como Hombres. Random forest

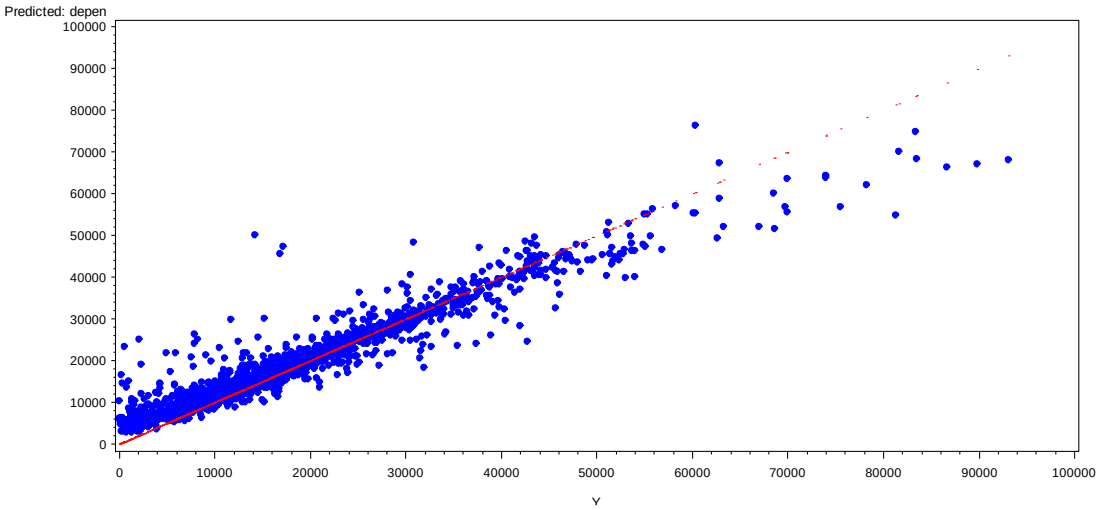
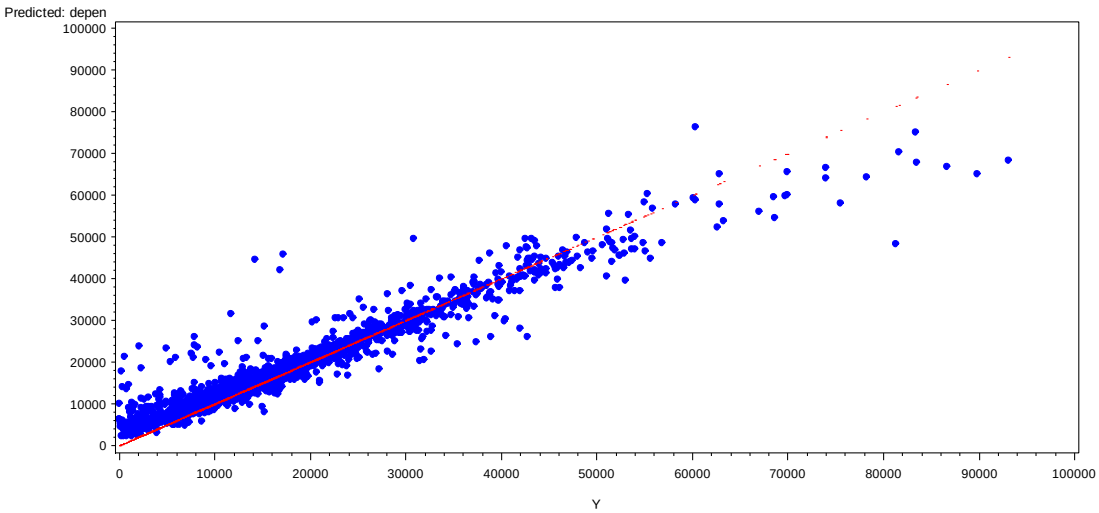


Ilustración 32. Predicción Salario Mujeres evaluadas como Hombres. Bagging.



II. Conclusiones y reflexión final. Un nuevo punto de partida.

Con este trabajo hemos podido probar diferentes algoritmos de Machine Learning para valorar cuál de ellos ayuda a predecir el salario a percibir por cada individuo de acuerdo con sus características personales, profesionales y del entorno laboral y familiar. Este estudio tiene la novedad de estimar ecuaciones de salarios con métodos diferentes a la clásica regresión lineal por mínimos cuadrados ordinarios usada en los estudios econométricos. De este modo se ha podido demostrar que existen otros modelos que se pueden ajustar bien a los datos y tienen una capacidad predictiva superior con menor error.

Este trabajo se ha centrado en los algoritmos y modelos expuestos, pero existe infinitud de modelos que se podrían probar e incluso demostrar que son mejores, o bien mejorar los existentes. Por ello se plantea para posibles trabajos futuros el uso de otros diferentes modelos y la mejora de estos.

Una de las limitaciones que se han encontrado es la falta de adaptabilidad de algunos de los modelos, en cuanto a capacidad predictiva, esto se debe a las características de los datos. En verdad la variable dependiente sigue una distribución mixta, pues toma valores igual a cero para aquellos individuos que quieren trabajar, pero no perciben salario (parados) y valores distintos de cero en caso contrario, ocupados. Esta limitación provocada por el sesgo muestral realizado en este estudio se podría solucionar en posteriores trabajos.

Como dato a destacar, la ecuación de salarios masculina a diferencia de la femenina si tiene en cuenta la variable número de hijos, mientras que la femenina no. ¿Por qué?, aquellos hombres que son padres tienen mayores incentivos a trabajar más horas y a crear un mayor vínculo con la empresa en búsqueda de la estabilidad laboral e ingresos recurrentes. Sin embargo, para las mujeres no es tenida en cuenta, porque el tipo de jornada influye más en la ecuación de salarios, la jornada reducida en las mujeres lleva implícito el hecho de tener o no tener hijos en búsqueda de la conciliación familiar.

Una de las preguntas que se plantean a raíz de este estudio es: ¿Qué es necesario hacer para hacer desaparecer discriminación salarial indirecta? Se deberían hacer tres cosas de forma simultánea:

1. Un endurecimiento de la ley antidiscriminación. Donde la ley obligaría a las empresas a eliminar la brecha salarial entre mujeres y hombres.
2. Una reducción de la segregación ocupacional de género e igual retribución del trabajo para mismo nivel de cualificación.
3. E igual distribución del tiempo de trabajo doméstico y de cuidados entre mujeres y hombres lo que implica plena corresponsabilidad. Si se consigue que los hombres concilien de igual forma que las mujeres se habrá conseguido un avance en este último factor.

III. Referencias bibliográficas.

Adler, Marina, y Lenz Karl. 2015. *Father Involvement in the Early Years: An International Comparison of Policy and Practice*. Policy Press.

Becker, Gary Stanley, y Gary S. Becker. 2009. *A Treatise on the Family*. Harvard university press.

Burnett, Simon B., Caroline J. Gatrell, Cary L. Cooper, y Paul Sparrow. 2013. *Fathers at work: A ghost in the organizational machine*. Gender, Work & Organization 20 (6): 632-46.

Calviño Martínez, A., 2018. *Apuntes en Técnicas y Metodología de la Minería de Datos*, Facultad de Estudios Estadísticos, Universidad Complutense de Madrid.

Conde-Ruiz, J Ignacio, e Ignacio Marra de Artíñano. 2016. *Gender Gaps in the Spanish Labor Market*, Estudios sobre la Economía Española -2016/32, , 103.

Cornejo, José Andrés Fernández, y Lorenzo Escot. 2018. *Brecha madre-padre en el uso de las medidas de conciliación y su efecto sobre las carreras profesionales de las madres*, junio, 167.

Cornejo, José Andrés Fernández, y Lorenzo Escot. 2013. *Panorama laboral 2013. análisis de algunas de las causas últimas de la desigualdad de género en el mercado laboral. implicaciones para las políticas de empleo,* septiembre.

Correll, Shelley J., Stephen Benard, y In Paik. 2007. *Getting a Job: Is There a Motherhood Penalty?*, American Journal of Sociology 112 (5): 1297-1339. https://doi.org/10.1086/511799.

García, R. Trabajo fin de grado: *Brecha Salarial de Género en España. Principales resultados después de la Crisis Económica del 2008.* Facultad de Ciencias Económicas y Empresariales, Universidad Complutense de Madrid, junio 2018.

Hobson, Barbara, y Susanne Fahlén. 2009. *Competing Scenarios for European Fathers: Applying Sen's Capabilities and Agency Framework to Work—Family Balance*, The ANNALS of the American Academy of Political and Social Science 624 (1): 214-33. https://doi.org/10.1177/0002716209334435.

INE, 2019. Encuesta de Estructura Salarial y Encuesta de Condiciones de Vida.

Kuhn, Max. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2018). *caret: Classification and Regression Training.* R package version 6.0-81. https://CRAN.R-project.org/package=caret

Liaw, A. and Wiener, M. (2002). *Classification and Regression by randomForest.* R News 2(3), 18--22.

País, El. 2016. *El cambio de las condiciones laborales y domésticas de los progenitores*. *El País*, 23 de junio de 2016, sec. Politica. https://elpais.com/elpais/2016/06/23/media/1466694667_389361.html.

Portela, J., 2019. *Apuntes sobre Técnicas de Machine Learning*, Facultad de Estudios Estadísticos, Universidad Complutense de Madrid.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ruzafa, I.,2018. Trabajo fin de grado: *Depuración de una base de datos,* Facultad de Estudios Estadísticos, Universidad Complutense de Madrid.

SAS® software (SAS Institute. 2012).

Sen, Amartya. 1990. *Development as capability expansion*. The community development reade*r*, 41-58.

Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer,New York. ISBN 0-387-95457-0.

IV. Anexos.

Anexo 1. Datos preliminares.

A continuación, se muestran los conjuntos de datos empleados para hacer los gráficos introductorios del trabajo que han ayudado a entender el panorama del mercado laboral en España.

Tabla 33. Tasa de Actividad, Paro y Ocupación en España (2006-2018).

	Tasa de Actividad (%)		Tasa de Paro (%)		Tasa de Ocupación (%)	
Año	**Hombres**	**Mujeres**	**Hombres**	**Mujeres**	**Hombres**	**Mujeres**
2006	69.22	48.45	6.35	11.34	93.65	88.66
2007	69.4	49.51	6.41	10.7	93.59	89.30
2008	69.54	50.93	10.05	12.84	89.95	87.16
2009	68.64	52.01	17.64	18.13	82.36	81.87
2010	68.15	52.7	19.57	20.22	80.43	79.78
2011	67.56	53.39	21.04	21.81	78.96	78.19
2012	67.1	53.98	24.58	25.03	75.42	74.97
2013	66.39	53.94	25.6	26.67	74.40	73.33
2014	65.83	53.67	23.6	25.43	76.40	74.57
2015	65.69	53.7	20.77	23.55	79.23	76.45
2016	65.13	53.64	18.12	21.38	81.88	78.62
2017	64.73	53.24	15.66	19.03	84.34	80.97
2018	64.55	53.06	13.72	17.02	86.28	82.98

Fuente: Encuesta Población Activa, INE.

Tabla 34. Evolución de los Salarios Brutos Anuales por sexo, el ratio mujer/hombre y la brecha salarial de género en España (2008-2018).

Año	**Mujeres**	**Hombres**	**Ambos sexos**	**GAP**	**Ratio Mujer/Hombre**	**Brecha**
2008	18,910.62 €	24,203.33 €	21,883.42 €	5,292.71 €	78.13%	21.87%
2009	19,502.02 €	25,001.05 €	22,511.47 €	5,499.03 €	78.00%	22.00%
2010	19,735.22 €	25,479.74 €	22,790.20 €	5,744.52 €	77.45%	22.55%
2011	19,767.59 €	25,667.89 €	22,899.35 €	5,900.30 €	77.01%	22.99%
2012	19,537.33 €	25,682.05 €	22,726.44 €	6,144.72 €	76.07%	23.93%
2013	19,514.58 €	25,675.17 €	22,697.86 €	6,160.59 €	76.01%	23.99%
2014	19,744.82 €	25,727.24 €	22,858.17 €	5,982.42 €	76.75%	23.25%
2015	20,051.58 €	25,992.76 €	23,106.30 €	5,941.18 €	77.14%	22.86%
2016	20,131.41 €	25,924.43 €	23,156.34 €	5,793.02 €	77.65%	22.35%
2017	20,607.85 €	26,391.84 €	23,646.50 €	5,783.99 €	78.08%	21.92%

Fuente: Encuesta Estructura Salarial, INE.

Anexo 2. Variables utilizadas para estimar la ecuación de salarios.

A continuación, se muestran las variables que se han seleccionado de los micro datos de la ECV de 2018 para estimar las ecuaciones de salarios.

Tabla 35. Descripción de las Variables utilizadas en el análisis de la descomposición de la Brecha Salarial.

Grado de urbanización			
Zona muy poblada	ZmuyPoblada	dicotómica	indica el grado de urbanización del lugar de residencia, tomando el valor 1 cuando se cumple el evento y 0 en caso contrario
Zona media	ZmedioPoblada	dicotómica	
Zona poco poblada	ZpocoPoblada	dicotómica	
Región			
Galicia	ES11		indica en que Comunidad Autónoma vive cada uno de los individuos
Principado de Asturias	ES12		
Cantabria	ES13		
País Vasco	ES21		
Comunidad Foral de Navarra	ES22		
La Rioja	ES23		
Aragón	ES24		
Comunidad de Madrid	ES30		
Castilla y León	ES41		
Castilla-La Mancha	ES42		
Extremadura	ES43		
Cataluña	ES51		
Comunidad Valenciana	ES52		
Illes Balears	ES53		
Andalucía	ES61		
Región de Murcia	ES62		
Ciudad Autónoma de Ceuta	ES63		
Ciudad Autónoma de Melilla	ES64		
Canarias	ES70		
Extra-Regio	ESZZ		

Edad			
Variable de tipo intervalo que indica la edad de cada individuo. Se ha calculado como la diferencia entre el año de la encuesta el año del nacimiento. Por ejemplo: 2018-1996=22			
Número de Hijos			
Variable de tipo intervalo que indica el número de hijos con edades entre 0 y 24 años inclusive conviviendo en los hogares y que son dependientes económicamente			
Hijos0_2		dicotómica	Además de la variable número de hijos se han creado las variables dicotómicas Hijos que toman el valor 1 cuando tienen hijos con edades dentro del rango y 0 en caso contrario
Hijos3_5		dicotómica	
Hijos6_12		dicotómica	
Mujer			
Mujer		1	Variable dicotómica. Esta variable se ha usado para filtrar la base de datos y extraer una para mujeres y otra para hombres.
Varón		0	
Casado o Conviviendo			
Cónyuge	casadooconviviendo	1	Variable dicotómica que toma el valor 1 cuando es cónyuge o 0 cuando es pareja de hecho
Pareja de hecho	casadooconviviendo	0	
Extranjero			
Sí		1	Variable dicotómica que indica si el individuo es de país de nacimiento extranjero o no
No (España)		0	
Estudiante			
Sí		1	Variable dicotómica, que toma el valor 1 si actualmente se está cursando algún tipo de estudios o el valor 0 si no es el caso
No		0	
Máximo nivel de estudios			
Sin estudios, analfabetos		0	Variable categórica que indica cual es el máximo nivel de estudios reglados terminados
Primaria	estudios de primaria	1	

Secundaria1	estudios de secundaria de primera etapa	2	
Secundaria2	estudios de secundaria de segunda etapa	3	
Superiores	estudios superiores (incluye FPII y Universidad)	4	

Experiencia en Estudios

Variable intervalo calculada como la diferencia entre el año actual y el año en que se terminó el máximo nivel de estudios terminados. Por ejemplo: 2018-2014= 4 años			
Situación Laboral			
Ocupado		dicotómica	Variables dicotómicas que toman el valor 1 cuando se produce el evento y 0 en caso contrario
Asalariado		dicotómica	
Autónomo		dicotómica	
Tiempo Parcial			
Variable dicotómica que indica si el individuo está asalariado a tiempo parcial o no			
Ocupación			
Ocu1_Direc	Dirección de empresas y de las Administraciones Públicas	dicotómica	Variables dicotómicas que toman el valor 1 cuando se produce el evento y 0 en caso contrario
Ocu2_Tecnicos	Técnicos profesionales científicos e intelectuales	dicotómica	
Ocu3_Tecapoyo	Técnicos y Profesionales de apoyo	dicotómica	
Ocu4_Administrativo	Empleados de tipo administrativo	dicotómica	
Ocu5_Servicios	Trabajadores de servicios de restauración, personales, protección y vendedores de comercio	dicotómica	

Ocu6_Agricultura	Trabajadores cualificados en la agricultura y en la pesca	dicotómica	
Ocu7_Cualificados	Artesanos y trabajadores cualificados de las industrias manufactureras, la construcción y la minería, excepto los operadores de instalaciones y maquinaria	dicotómica	
Ocu8_Operadores	Operadores de instalaciones y maquinaria, y montadores	dicotómica	
Ocu9_NoCualifica	Trabajadores no cualificados	dicotómica	
Horas de Trabajo en Total			
Variable de tipo intervalo que muestra el número de horas normalmente trabajadas a la semana entre todos los trabajos (principal y secundarios).			
Actividad del establecimiento del que depende o dependía laboralmente			Variables dicotómicas que toman el valor 1 cuando se produce el evento y 0 en caso contrario
Act1_Agric	Agricultura, ganadería, silvicultura y pesca	dicotómica	
Act2_Ind	Industrias extractivas, manufacturera y de Suministro de energía eléctrica, gas, agua y gestión de residuos.	Dicotómica	
Act3_Construc	Construcción	dicotómica	
Act4_Comerc	Comercio, reparación de vehículos de motor	dicotómica	
Act5_Hostel	Hostelería	dicotómica	
Act6_Trans	Transporte y almacenamiento, Información y comunicaciones	dicotómica	
Act7_Financ	Actividades financieras y de seguros	dicotómica	

Act8_ServEmpres	Actividades inmobiliarias, Actividades profesionales, científicas y técnicas y Actividades administrativas y servicios auxiliares	dicotómica	
Act9_AAPP	Administración pública y defensa. Seguridad Social	dicotómica	
Act10_Educ	Educación	dicotómica	
Act11_Sanidad	Actividades sanitarias y de servicios sociales	dicotómica	
Act12_OtrosServ	Otros servicios, Actividades artísticas, recreativas, Hogares como empleadores de personal doméstico, Organismos extraterritoriales y no consta	dicotómica	
Tamaño de la Empresa			
Tamanoempre_1_10	1 a 10 personas	dicotómica	Número de personas que trabajan en la empresa
Tamanoempre_11_19	11 a 19 personas	dicotómica	
Tamanoempre_20_49	20 a 49 personas	dicotómica	
Tamanoempre_50omas	50 personas o más	dicotómica	
Temporal			
Variable dicotómica que toma el valor 1 si el contrato laboral es de duración determinada y 0 en caso contrario			
Supervisor			
Variable dicotómica que toma el valor 1 si en su puesto de trabajo supervisa o coordina el trabajo de algún empleado de la empresa u organismo y 0 en caso contrario			
Experiencia			
Variable de tipo intervalo que es calculada como la diferencia entre la edad actual y la edad en la que empezó a trabajar. Por ejemplo: 45-23=22			

Estado general de salud			
Muy bueno		dicotómica	Variables dicotómicas que toman el valor 1 cuando se produce el evento y 0 en caso contrario
Bueno		dicotómica	
Regular		dicotómica	
Malo		dicotómica	
Muy Malo		dicotómica	
Limitaciones de Salud			
Gravemente limitado		dicotómica	Variables dicotómicas que toman el valor 1 cuando se produce el evento y 0 en caso contrario
Limitado, pero no gravemente		dicotómica	
Nada limitado		dicotómica	
Salario Neto			
Variable objetivo. Indica el salario neto mensual actual. Es la variable dependiente de la Ecuación de salarios, es de tipo intervalo			
Renta Familiar Neta			
Es la Renta Familiar Neta del año anterior, es decir, 2017. Se ha recodificado de tal manera que los valores negativos sean ceros.			
Renta Familiar Neta del Resto de la Familia			
Es la diferencia entre la Renta Familiar Neta y el Salario Bruto. Se ha recodificado de tal manera que los valores negativos sean ceros.			
Experiencia_2			
Transformación cuadrática de la variable Experiencia			
Edad_2			
Transformación cuadrática de la variable Edad			

Tabla 36. Análisis descriptivo individual de cada una de las variables de tipo intervalo para conjunto de datos de mujeres.

Variable	Rol	Media	Desviación estándar	No ausente	Ausente	Mínimo	Mediana	Máximo	Asimetría	Curtosis
Edad	INPUT	44.55164	10.56289	5558	0	18	45	64	-0.25057	-0.76452
Edad_2	INPUT	2096.403	921.1951	5558	0	324	2025	4096	0.164561	-0.88344
Experiencia	INPUT	23.75014	11.66325	5547	11	0	24	53	-0.03776	-0.73461
Experiencia_2	INPUT	698.6902	570.0562	5558	0	0	576	2809	0.836917	-0.02929
Experienciaestudios	INPUT	23.33429	12.50135	5558	0	0	23	54	-0.03258	-0.8932
HorastrabajoTotal	INPUT	36.0968	9.843328	5558	0	2	40	90	-0.56408	2.221093
RentaFamiliarNeta	INPUT	3914426	2446292	5558	0	0	3470980	29252100	2.217856	11.3083
RentaFamiliarneta_RestoFamilia	INPUT	2505256	2125525	5558	0	0	2136360	27396180	2.943699	19.58678
SalarioNeto	TARGET	17512.76	15359.23	5558	0	0	14673.8	229012.6	1.908091	10.05302

Tabla 37. Análisis descriptivo individual de cada una de las variables de tipo intervalo para conjunto de datos de hombres.

Variable	Rol	Media	Desviación estándar	No ausente	Ausente	Mínimo	Mediana	Máximo	Asimetría	Curtosis
Edad	INPUT	44.5721	10.89721	6207	0	18	45	64	-0.28037	-0.76268
Edad_2	INPUT	2105.402	946.8043	6207	0	324	2025	4096	0.141222	-0.93599
Experiencia	INPUT	25.30717	12.00803	6192	15	0	26	56	-0.16905	-0.7041
Experiencia_2	INPUT	782.7263	606.0562	6207	0	0	676	3136	0.696553	-0.3093
Experienciaestudios	INPUT	24.38167	12.82366	6207	0	0	25	54	-0.13736	-0.9154
HorastrabajoTotal	INPUT	41.41308	7.974893	6207	0	3	40	99	0.634311	5.919398
RentaFamiliarNeta	INPUT	3843523	2387729	6207	0	0	3377730	25898080	2.210018	10.39793
RentaFamiliarneta_RestoFamilia	INPUT	2142651	1957169	6207	0	0	1732810	23577117	2.555771	13.9056
SalarioNeto	TARGET	21848.31	20866.26	6207	0	0	18745.7	398312.6	3.936809	40.24196

Tabla 38. Análisis descriptivo individual de cada una de las variables de clase de datos de mujeres.

Rol de los datos	Nombre de la variable	Rol	Número de niveles	Ausente	Moda	Porcentaje moda
TRAIN	Act10_Educ	INPUT	3	1	0	88.27
TRAIN	Act11_Sanid	INPUT	3	1	0	83.72
TRAIN	Act12_OtrosSer	INPUT	3	1	0	89.94
TRAIN	Act1_Agr	INPUT	3	1	0	97.84
TRAIN	Act2_Ind	INPUT	3	1	0	90.72
TRAIN	Act3_Constr	INPUT	3	1	0	98.78
TRAIN	Act4_Comer	INPUT	3	1	0	84.78
TRAIN	Act5_Hostel	INPUT	3	1	0	92.17
TRAIN	Act6_Trans	INPUT	3	1	0	95.16
TRAIN	Act7_Finan	INPUT	3	1	0	96.96
TRAIN	Act8_ServEmpr	INPUT	3	1	0	88.86
TRAIN	Act9_AAPP	INPUT	3	1	0	92.61
TRAIN	Asalariado	INPUT	2	0	1	90.25
TRAIN	Aut_nomo	INPUT	2	0	0	90.25
TRAIN	CasadooConviviendo	INPUT	3	1850	1	59.36
TRAIN	Comunidad_Aut_noma	INPUT	19	0	ES51	23.35
TRAIN	Estudiante	INPUT	2	0	0	96.19
TRAIN	Extranjero	INPUT	2	0	1	71.79
TRAIN	LimitacionSaludGrave	INPUT	2	0	0	99.06
TRAIN	LimitacionSaludLeve	INPUT	2	0	0	90.36
TRAIN	LimitacionSaludNinguna	INPUT	2	0	1	89.42
TRAIN	MaximoNivelEstudios	INPUT	5	0	4	51.66
TRAIN	NumHijos	INPUT	7	131	0	45.23
TRAIN	Ocu1_Direc	INPUT	2	0	0	97.48
TRAIN	Ocu2_Tecnicos	INPUT	2	0	0	78.09
TRAIN	Ocu3_Tecapoyo	INPUT	2	0	0	87.26
TRAIN	Ocu4_Administrativo	INPUT	2	0	0	80.59
TRAIN	Ocu5_Servicios	INPUT	2	0	0	77.19
TRAIN	Ocu6_Agricultura	INPUT	2	0	0	99.15
TRAIN	Ocu7_Cualificado	INPUT	2	0	0	96.98
TRAIN	Ocu8_Operador	INPUT	2	0	0	98.38
TRAIN	Ocu9_NoCualif	INPUT	2	0	0	84.89
TRAIN	SaludBuena	INPUT	2	0	1	60.58
TRAIN	SaludMala	INPUT	2	0	0	98.24
TRAIN	SaludMuyBuena	INPUT	2	0	0	74.22
TRAIN	SaludMuyMala	INPUT	2	0	0	99.75
TRAIN	SaludRegular	INPUT	2	0	0	88.38
TRAIN	Supervisor	INPUT	3	570	0	74.83
TRAIN	Tamanoempresa_11_19	INPUT	3	134	0	86.04
TRAIN	Tamanoempresa_1_10	INPUT	3	134	0	59.86
TRAIN	Tamanoempresa_20_49	INPUT	3	134	0	83.52
TRAIN	Tamanoempresa_50omas	INPUT	3	134	0	65.92
TRAIN	Temporal	INPUT	3	637	0	65.89
TRAIN	Tiempo_Parcial	INPUT	2	0	0	79.74
TRAIN	ZMedioPoblada	INPUT	2	0	0	77.06
TRAIN	ZMuyPoblada	INPUT	2	0	1	53.33
TRAIN	ZPocoPoblada	INPUT	2	0	0	76.27

Tabla 39. Análisis descriptivo individual de cada una de las variables de clase de datos de hombres.

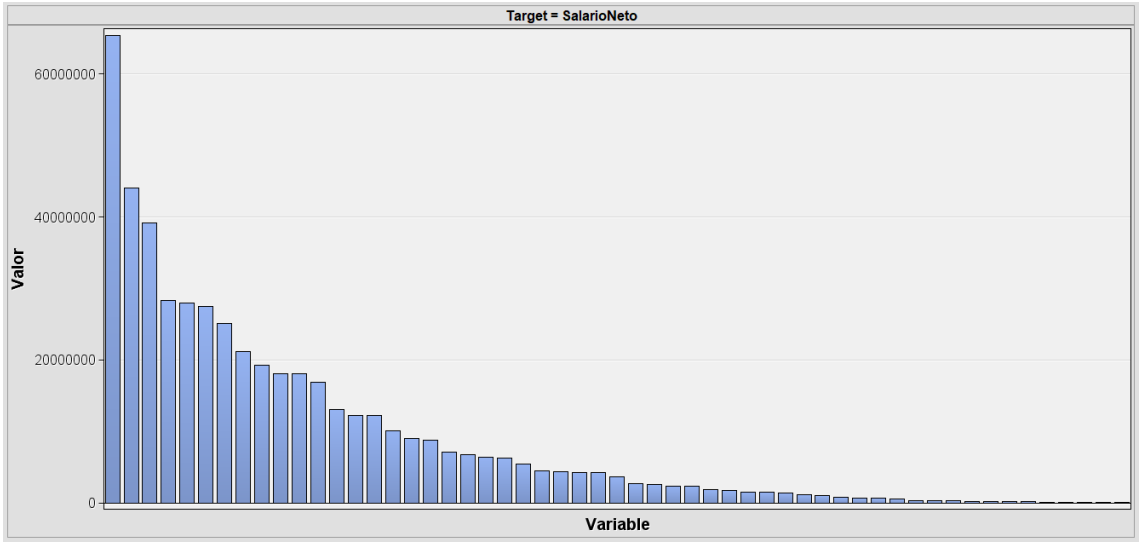
Rol de los datos	Nombre de la variable	Rol	Número de niveles	Ausente	Moda	Porcentaje moda
TRAIN	Act10_Educ	INPUT	2	0	0	95.05
TRAIN	Act11_Sanid	INPUT	2	0	0	96.15
TRAIN	Act12_OtrosSer	INPUT	2	0	0	96.54
TRAIN	Act1_Agr	INPUT	2	0	0	94.78
TRAIN	Act2_Ind	INPUT	2	0	0	77.86
TRAIN	Act3_Constr	INPUT	2	0	0	89.16
TRAIN	Act4_Comer	INPUT	2	0	0	86.72
TRAIN	Act5_Hostel	INPUT	2	0	0	94.33
TRAIN	Act6_Trans	INPUT	2	0	0	89.14
TRAIN	Act7_Finan	INPUT	2	0	0	97.81
TRAIN	Act8_ServEmpr	INPUT	2	0	0	91.36
TRAIN	Act9_AAPP	INPUT	2	0	0	91.09
TRAIN	Asalariado	INPUT	2	0	1	85.16
TRAIN	Aut_nomo	INPUT	2	0	0	85.16
TRAIN	CasadooConviviendo	INPUT	3	1891	1	62.48
TRAIN	Comunidad_Aut_noma	INPUT	19	0	ES51	22.25
TRAIN	Estudiante	INPUT	2	0	0	97.05
TRAIN	Extranjero	INPUT	3	2	1	72.87
TRAIN	LimitacionSaludGrave	INPUT	2	0	0	99.29
TRAIN	LimitacionSaludLeve	INPUT	2	0	0	92.85
TRAIN	LimitacionSaludNinguna	INPUT	2	0	1	92.14
TRAIN	MaximoNivelEstudios	INPUT	5	0	4	39.31
TRAIN	NumHijos	INPUT	8	126	0	47.19
TRAIN	Ocul_Direc	INPUT	3	1	0	94.35
TRAIN	Ocu2_Tecnicos	INPUT	3	1	0	86.31
TRAIN	Ocu3_Tecapoyo	INPUT	3	1	0	87.82
TRAIN	Ocu4_Administrativo	INPUT	3	1	0	92.46
TRAIN	Ocu5_Servicios	INPUT	3	1	0	85.07
TRAIN	Ocu6_Agricultura	INPUT	3	1	0	96.70
TRAIN	Ocu7_Cualificado	INPUT	3	1	0	80.28
TRAIN	Ocu8_Operador	INPUT	3	1	0	88.34
TRAIN	Ocu9_NoCualif	INPUT	3	1	0	88.56
TRAIN	SaludBuena	INPUT	2	0	1	61.62
TRAIN	SaludMala	INPUT	2	0	0	98.87
TRAIN	SaludMuyMala	INPUT	2	0	0	99.87
TRAIN	SaludRegular	INPUT	2	0	0	89.32
TRAIN	Supervisor	INPUT	3	995	0	63.56
TRAIN	Tamanoempresa_11_19	INPUT	3	171	0	83.91
TRAIN	Tamanoempresa_1_10	INPUT	3	171	0	62.16
TRAIN	Tamanoempresa_20_49	INPUT	3	171	0	84.05
TRAIN	Tamanoempresa_50omas	INPUT	3	171	0	64.23
TRAIN	Temporal	INPUT	3	1015	0	63.96
TRAIN	Tiempo_Parcial	INPUT	2	0	0	95.62
TRAIN	ZMedioPoblada	INPUT	2	0	0	76.48
TRAIN	ZMuyPoblada	INPUT	2	0	1	50.04
TRAIN	ZPocoPoblada	INPUT	2	0	0	73.56

Anexo 3. Variables más importantes para la estimación de la ecuación de salarios.

En este anexo se exponen los gráficos de importancia de la variable para el conjunto de datos de mujeres y de hombres. Estos gráficos asignan un valor, importancia, a cada una de las variables input con respecto a la variable output.

El primero de los gráficos muestra de forma compacta la relevancia de cada una de las variables ordenadas de mayor a menor. El segundo de los gráficos es una ampliación del gráfico inicial para que se pueda ver que barra representa a cada variable, dado que la salida de SAS no lo puede representar todo a la vez porque tenemos un gran número de variables.

Ilustración 33. Importancia de la Variable para conjunto de datos mujeres.



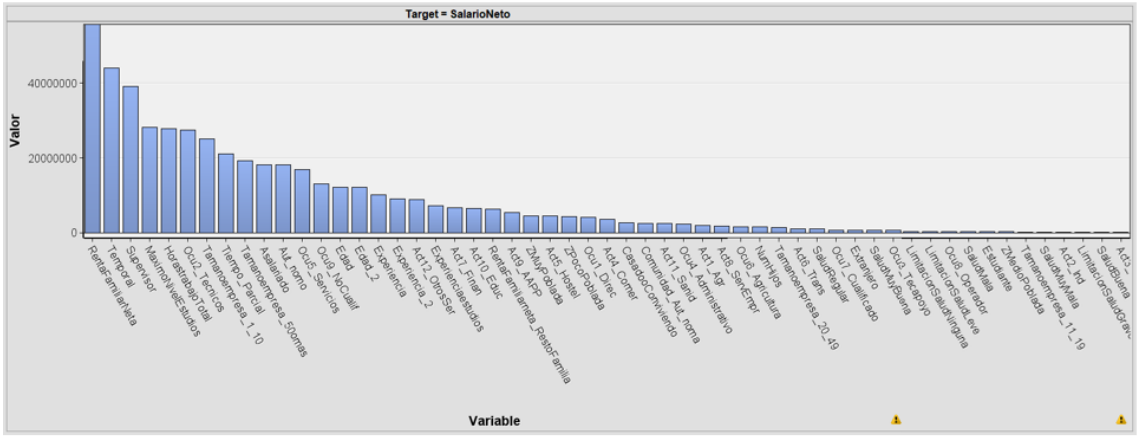


Ilustración 34. Importancia de la Variable para conjunto de datos hombres.

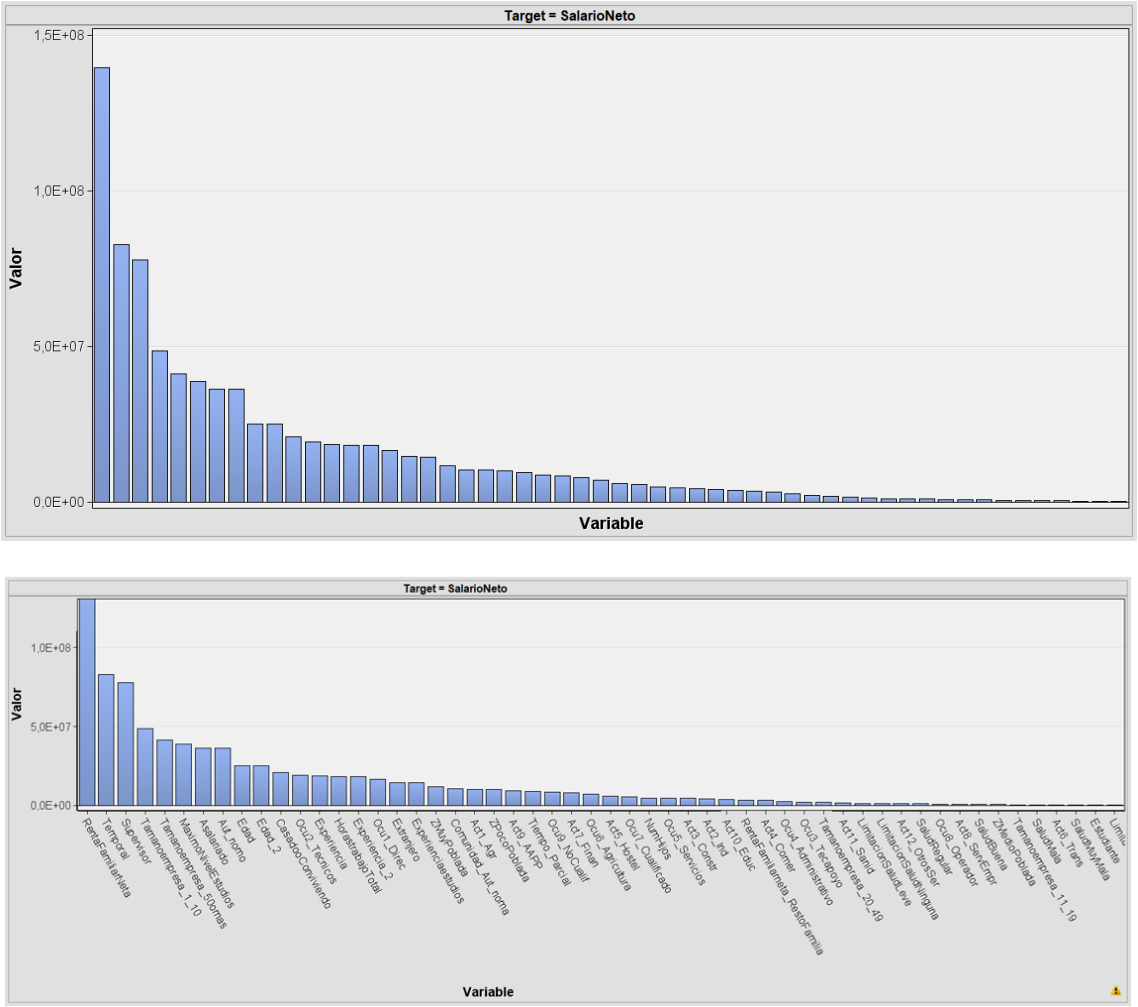


Tabla 40. Reagrupación de las variables categóricas: máximo nivel de estudios, comunidad autónoma y número de hijos.

Variable	Recategorización	Niveles iniciales
Comunidad Autónoma	0	ES61 ES62 ES63 ES64
	1	ES42 ES43
	2	ES11 ES12 ES13 ES41 ES52 ES70
	3	ES23 ES24
	4	ES21 ES30
	5	ES22 ES51
	6	ES53
Máximo nivel de estudios	0	0 1
	1	2
	2	3
	3	4
Número de hijos	0	0, 1, 2 hijos
	1	3, 4 hijos
	2	5, 8 hijos

Anexo 4. Código usado en SAS® software y en R®.

Todo el código usado en SAS ®Software y en R Studio, así como los datos se puede consultar en la siguiente dirección web a una carpeta compartida.

https://bit.ly/2MUuT8A

Esta carpeta compartida contiene los scripts ejecutados en ambos programas para todo el caso de estudio que se ha expuesto en este trabajo de investigación, así como las fuentes de datos generadas y macros usadas.