Texto

Descripción generada automáticamente

**Trabajo Final de la Diplomatura en Ciencias Sociales Computacionales y Humanidades Digitales**

Módulos 3 y 4

*Universidad Nacional de San Martín*

Gonzalo Daniel Azuaga

1. **Introducción**

El presente trabajo tiene como fin tratar de predecir el ingreso de la ocupación principal de los individuos utilizando diferentes variables relevadas en la Encuesta Permanente de Hogares correspondiente (EPH) al II trimestre del año 2020. Dentro del *dataset* seleccionado, sólo están registradas las personas que declararon haber trabajado al menos una hora durante la semana anterior al operativo, es decir ocupadas, y se han elegido algunas variables para el análisis. La base cuenta con 12803 filas, cada una de ellas representa a un individuo y 14 variables. A continuación, se renombran las variables y se presenta una tabla con las definiciones de cada una de ellas (Cuadro 1).

| **Cuadro 1. Definición de variables** | |
| --- | --- |
| Variable | Descripción |
| Codusu | número de identificación que permite relacionar una vivienda con hogares y personas |
| Nro\_hogar | refiere al hogar dentro de la vivienda. Puede haber más de un hogar en una vivienda |
| Nivel\_ed | máximo nivel educativo alcanzado |
| Relacion | relación de parentesco con el jefe/a de hogar |
| Sexo | sexo |
| Edad | edad en años cumplidos |
| Estado | estado conyugal |
| Tipo | sector del establecimiento donde trabaja (público, privado) |
| Categoria\_ocup | categoría ocupacional |
| Intensidad | intensidad de la ocupación (ocupado pleno, sobreocupado, subocupado) |
| Total\_horas | cantidad de horas que trabajó la semana anterior en la ocupación principal |
| Categoria | carácter de la ocupación principal (basado en el CNO) |
| Calificacion | calificación de la ocupación principal (basado en el CNO) |
| Ingreso | Ingreso de la ocupación principal |
| **Fuente: EPH 2º Trimestre 2020.** | |

El trabajo se estructura de la siguiente manera: en la primera sección se realiza un análisis exploratorio del *dataset*, de forma gráfica y descriptiva, con la intención de analizar, describir, resumir y visualizar la naturaleza de los datos. En la segunda sección, se realiza la tarea de modelización. Para ello, en primer lugar, se define el conjunto de *training* y *testing*, luego se realiza un preprocesado de los datos (análisis de la varianza, estandarización, creación de variables *dummies*, selección de predictores más importantes, entre otros). En segundo lugar, se entrenan diferentes modelos para evaluar el de mejor performance y se interpreta su output final en función del problema de estudio abordado.

1. **Análisis exploratorio (EDA)**

| **Cuadro 2. Análisis exploratorio de variables categóricas.** | | | | |
| --- | --- | --- | --- | --- |
| Tipo | Variable | N\_únicos | Espacios\_vacíos | NA\_s |
| character | Codusu | 8338 | 0 | 0 |
| character | Nivel\_ed | 7 | 0 | 0 |
| character | Relacion | 10 | 0 | 0 |
| character | Sexo | 2 | 0 | 0 |
| character | Estado | 5 | 0 | 0 |
| character | Tipo | 4 | 0 | 0 |
| character | Categoria\_ocup | 4 | 0 | 0 |
| character | Intensidad | 4 | 0 | 0 |
| character | Categoria | 51 | 0 | 0 |
| character | Calificacion | 7 | 0 | 0 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | |

| **Cuadro 2.1. Análisis exploratorio de variables cuantitativas.** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tipo | Variable | NA\_s | Media | Desvío\_std | Mín | Q1 | Mediana | Q3 | Máx | Hist |
| numeric | Nro\_hogar | 0 | 1,06 | 1,22 | 1 | 1 | 1 | 1 | 61 | ▇▁▁▁▁ |
| numeric | Edad | 0 | 41,44 | 12,42 | 14 | 31 | 41 | 51 | 82 | ▃▇▇▃▁ |
| numeric | Total\_horas | 0 | 26,70 | 22,12 | 0 | 8 | 28 | 40 | 999 | ▇▁▁▁▁ |
| numeric | Ingreso | 0 | 23757,93 | 24437,94 | -9 | 6000 | 20000 | 35000 | 545000 | ▇▁▁▁▁ |
| **Fuente: EPH 2º Trimestre 2020.** | | | | | | | | | | |

A partir de los Cuadros resumen (2 y 2.1) generados con la función *skim* del paquete *skirm* se observa que, en principio, el *dataset* no cuenta con missing values (NA’s). Por otro lado, llama la atención el valor mínimo de la variable Ingreso (-9) y el máximo de la variable Total\_Horas (999). En el *Diseño de Registro y Estructura de la base de la EPH* se observa que el valor -9 corresponde a la no respuesta, mientras que el 999 indica la categoría de “No sabe/no responde”. Existen 1446 casos para el primero y sólo 1 para el segundo en el *dataset*. Por otro lado, 712 individuos reportaron un monto de ingreso de la ocupación principal de 0. Para el análisis exploratorio no se incluirán los casos de no respuesta de ambas variables, si recibirán un tratamiento particular al momento de la modelización.

* 1. *Características principales de las personas que no declararon ingreso de su ocupación principal*

A continuación, se presenta un breve análisis de las principales características de las personas que no declararon sus ingresos de sus ocupaciones principales, es decir que dentro del *dataset* se encuentran codificadas con -9 en la variable Ingreso.

En primer lugar, aproximadamente el 46% de las personas se reconocen como jefe/a del hogar (Cuadro 3), en segundo se percibe que aquellas personas con mayor nivel educativo tienden, en promedio, a no declarar su ingreso, casi el 62% de las personas cuenta con secundaria o universitario completo (Cuadro 4). En tercero, el 70,47% resultan empleados u obreros (Cuadro 5) mientras que aproximadamente la mitad de ellos (el 48,13%) son ocupados plenos (Cuadro 6). Por último, se observa una tendencia a que los varones sean más reticentes a declarar su ingreso en relación con las mujeres (Cuadro 6).

| **Cuadro 3. Relación con el Jefe/a del hogar** | | |
| --- | --- | --- |
| Relación | N | Freq |
| Jefe/a | 664 | 45,92% |
| Hijo/a Hijastro/a | 348 | 24,07% |
| Conyuge / Pareja | 328 | 22,68% |
| Yerno/Nuera | 35 | 2,42% |
| Hermano/a | 22 | 1,52% |
| Otros Familiares | 18 | 1,24% |
| Nieto/a | 13 | 0,9% |
| No familiares | 11 | 0,76% |
| Madre/Padre | 7 | 0,48% |
| **Fuente: EPH 2º Trimestre 2020.** | | |

| **Cuadro 4. Nivel educativo** | | |
| --- | --- | --- |
| Nivel\_ed | N | Freq |
| Superior universitaria completa | 448 | 30,98% |
| Secundaria completa | 444 | 30,71% |
| Superior universitaria incompleta | 203 | 14,04% |
| Secundaria incompleta | 192 | 13,28% |
| Primaria completa | 118 | 8,16% |
| Primaria incompleta (incluye educación especial) | 35 | 2,42% |
| Sin instrucción | 6 | 0,41% |
| **Fuente: EPH 2º Trimestre 2020.** | | |

| **Cuadro 5. Categoría ocupacional** | | |
| --- | --- | --- |
| Categoria\_ocup | N | Freq |
| Obrero o empleado | 1019 | 70,47% |
| Cuenta propia | 362 | 25,03% |
| Patrón | 65 | 4,5% |
| **Fuente: EPH 2º Trimestre 2020.** | | |

| **Cuadro 6. Intensidad de la ocupación** | | |
| --- | --- | --- |
| Intensidad | N | Freq |
| Ocupado pleno | 696 | 48,13% |
| Sobreocupado | 312 | 21,58% |
| Ocupado que no trabajo en la semana | 294 | 20,33% |
| Subocupado por insuficiencia horaria | 144 | 9,96% |
| **Fuente: EPH 2º Trimestre 2020.** | | |

| **Cuadro 7. Sexo** | | |
| --- | --- | --- |
| Sexo | N | Freq |
| Varón | 854 | 59,06% |
| Mujer | 592 | 40,94% |
| **Fuente: EPH 2º Trimestre 2020.** | | |
|  | | |

* 1. *Análisis de las variables cuantitativas*

| **Cuadro 8. Medidas descriptivas: tendencia central, dispersión y asimetría** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Media | Desvío\_std | Mín | Q1 | Mediana | Q3 | Máx | CV | Asimetría |
| Total\_horas | 26,48 | 20,37 | 0 | 7 | 27 | 40 | 168 | 76,91 | -0,08 |
| Ingreso | 26786,35 | 24334,39 | 0 | 10000 | 23000 | 36000 | 545000 | 90,85 | 0,47 |
| Edad | 41,32 | 12,36 | 14 | 31 | 41 | 50 | 82 | 29,91 | 0,08 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | | | | | | |

Gráfico, Histograma

Descripción generada automáticamente con confianza media

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

A partir del Cuadro 8 y de los Gráficos 1 y 2, es posible observar cómo se distribuyen las variables continuas del *dataset*. En promedio, las personas relevadas en la encuesta (ocupadas) poseen un ingreso de su ocupación principal de $26786,35, una edad media de 41 años y han trabajado en la última semana 27 horas aproximadamente. Con respecto a la dispersión se aprecia, a partir de la estimación del Coeficiente de Variación (CV) y de los gráficos (histogramas y diagramas de caja y bigote), que la variable “Edad” posee una menor variabilidad en torno a su media en comparación al resto de las variables (presenta el menor CV). Por otro lado, dicha variable resulta ser la más simétrica de las 3, aspecto que se observa, de forma analítica, al comparar el Coeficiente de Asimetría de Pearson estimado (el mismo es bastante cercano a 0) y, a su vez, de forma visual, a través de los histogramas (la distribución de dicha variable tiende asimilarse a una campana) y en el diagrama de caja y bigote se observa que el largo del bigote inferior y superior es bastante similar.

La distribución de la variable “Ingreso” resulta ser la más asimétrica (registra una marcada asimetría positiva), en tanto posee un Coeficiente de Asimetría de Pearson mayor a 0 (0,47) y que resulta superior al resto de las distribuciones, lo que indica que la gran mayoría de las personas relevadas declararon ingresos menores a la media (subpromediales). En efecto, el 75% de las personas registraron ingresos de la ocupación principal de hasta $36000 (Cuadro 8). Este aspecto también se corrobora al observar tanto el histograma como el diagrama de caja y bigote. La variable “Total\_horas” presenta una leve asimetría negativa, es decir que, en principio, la mayoría de las personas trabajaron durante la última semana más que la media.

Un aspecto que se destaca del análisis exploratorio resulta la existencia de datos atípicos en las variables “Ingreso” y “Total\_horas”. Tal como se destacó con anterioridad, el 75% de las personas registraron ingresos de hasta $36000, mientras que el restante 25% posee ingresos que oscilan entre dicho valor y $545000, lo que denota la gran variabilidad que existe en este último segmento de la muestra. Para los análisis y visualizaciones posteriores se aplicará el logaritmo en base 10 a dicha variable. Dicha transformación puede utilizarse cuando existen grandes disparidades en una variable, porque permite que los valores pequeños tengan espacio para diferenciarse, y a la vez que los muy grandes no aparezcan tan alejados, logrando, por lo general, visualizaciones más compactas y legibles.

Con respecto a la variable “Total\_horas”, el 75% de las personas relevadas trabajó en la última semana hasta 40 horas, el segmento restante fluctuó entre 40 y 168 horas. Este último valor representaría a una persona que trabajó en promedio 24 horas diarias durante 7 días. Dado que este dato puede resultar anómalo, es posible observar el carácter de la ocupación principal de las personas basado en el Clasificador Nacional de Ocupaciones (CNO), tal como se aprecia en el Cuadro 9. Existen 2 personas que declararon haber trabajado 168 horas en la última semana, las mismas se desempeñan en servicios policiales y en cuidado y atención de personas, actividades en las cuáles pueden darse situaciones en donde se trabajé en formato de guardia.

| **Cuadro 9. Categoría ocupacional de las 10 personas que más trabajaron en la semana anterior** | |
| --- | --- |
| Categoría | Total\_horas |
| Ocupaciones de servicios policiales | 168 |
| Ocupaciones del cuidado y la atención de personas | 168 |
| Ocupaciones del transporte | 126 |
| Ocupaciones de la comercialización directa | 119 |
| Ocupaciones del cuidado y la atención de personas | 112 |
| Ocupaciones de la producción ganadera | 112 |
| Ocupaciones de la comercialización directa | 112 |
| Ocupaciones de la comercialización directa | 105 |
| Directivos de pequeñas y microempresas | 101 |
| Directivos de medianas empresas privadas productoras de bienes y servicios | 98 |
| **Fuente: EPH 2º Trimestre 2020.** | |

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

| **Cuadro 10. Test de correlaciones** | | | | |
| --- | --- | --- | --- | --- |
| Variables | Estimado | Valor\_p | Lim. inferior | Lim.superior |
| Ingreso - Total\_horas | 0.18 | 0.000000000 | 0.16 | 0.20 |
| Ingreso - Edad | 0.15 | 0.000000000 | 0.14 | 0.17 |
| Edad - Total\_horas | -0.06 | 0.000000004 | -0.07 | -0.04 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | |

Los análisis de correlación indican que las 3 relaciones entre pares de las variables continuas del *dataset* son significativas, los valores p asociados a cada relación son menores a un nivel de significancia del 5% (Cuadro 10). De dichas relaciones la correspondiente a Ingreso y Total\_horas presenta mayor intensidad (lineal positiva), aunque cabe destacar que la misma es bastante leve (0,18 es el Coeficiente de Correlacion estimado). A partir del Gráfico 3 se observa que la gran mayoría de las personas trabajaron hasta 50 horas semanales, pero se percibe una gran variabilidad en torno al logaritmo del ingreso percibido en ese segmento. A su vez, es posible registrar los datos extremos de 168 horas trabajadas identificadas en el Cuadro 8.

Con respecto a las otras correlaciones, se observa que la Edad presenta una menor intensidad en su relación con el ingreso, y que la misma con las horas trabajadas es negativa, lo que indicaría que a medida que las personas crecen tienden a trabajar menos horas. Aunque cabe destacar que dicha correlación es de muy baja intensidad.

Para complementar el análisis de correlación se estimará la *Variance Inflation Factors* (VIF) para corroborar la posible existencia o no de multicolinealidad, según la literatura un VIF mayor a 5 o a 10 estaría indicando la presencia de multicolinealidad. La VIF para el predictor *j* se estima a partir de la siguiente fórmula:

=

Siendo: el Coeficiente de Determinación obtenido de la regresión del predictor *j* con el resto de los predictores. En este caso, al tener sólo 2 predictores cuantitativos continuos se obtendrá un único VIF.

| **Cuadro 11. Variance Inflation Factors** | |
| --- | --- |
| Variables | VIF |
| Total\_horas | 1,0031 |
| Edad | 1,0031 |
| **Fuente: EPH 2º Trimestre 2020.** | |

Los valores de VIF obtenidos en el Cuadro 11, sumado a la Matríz y Test de Correlación estimados, indicarían que no habría problemas de colinealidad y multicolinealidad en el *dataset.*

* 1. *Análisis de las variables categóricas*

Para una mejor visualización y posterior modelización se recodificaron las variables de Nivel\_ed, Relación, Estado, Intensidad, Tipo y Categoría (en Rmd adjuntado se explicita cómo fue dicha recodificación).

A continuación, se presenta cómo se distribuye el ingreso de la ocupación principal en función de todas las variables categóricas, de forma visual (diagramas de caja y bigote y densidad) y luego de forma analítica estimando distintas medidas de tendencia central, dispersión y asimetría.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamenteGráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamenteGráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

| **Cuadro 12. Distribución del ingreso de la ocupación principal según Sexo** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sexo | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Varón | 6318 | 55,64 % | 29372,35 | 25000 | 26998,10 | 91,92 | 0,49 |
| Mujer | 5038 | 44,36 % | 23543,33 | 20000 | 20047,23 | 85,15 | 0,53 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | | | | |

| **Cuadro 13. Distribución del ingreso de la ocupación principal según Nivel Educativo.** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Nivel\_ed | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Superior | 2809 | 24,74% | 39442,84 | 35000 | 31648,63 | 80,24 | 0,42 |
| Secundaria | 4939 | 43,49% | 25586,95 | 22800 | 21369,89 | 83,52 | 0,39 |
| Primaria | 3263 | 28,73% | 19009,37 | 15000 | 16441,15 | 86,49 | 0,73 |
| Sin instrucción | 345 | 3,04% | 14461,74 | 10000 | 14037,70 | 97,07 | 0,95 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | | | | |

| **Cuadro 14. Distribución del ingreso de la ocupación principal según la calificación del puesto de trabajo.** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Calificación | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Profesionales | 963 | 8,54% | 52438,71 | 45000 | 45976,63 | 87,68 | 0,49 |
| Técnicos | 2169 | 19,24% | 32334,30 | 30000 | 20640,23 | 63,83 | 0,34 |
| Operativos | 5865 | 52,04% | 24792,50 | 22000 | 19535,31 | 78,80 | 0,43 |
| No calificados | 2274 | 20,18% | 15347,40 | 12000 | 12656,16 | 82,46 | 0,79 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | | | | |

| **Cuadro 15. Distribución del ingreso de la ocupación principal según el Estado Conyugal.** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Estado | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Casado | 3656 | 32,19% | 32470,92 | 30000 | 29155,42 | 89,79 | 0,25 |
| Unido | 3068 | 27,02% | 26340,15 | 23000 | 20629,93 | 78,32 | 0,49 |
| Separado/a | 1072 | 9,44% | 26295,58 | 22000 | 22663,27 | 86,19 | 0,57 |
| Viudo/a | 196 | 1,73% | 22757,91 | 17000 | 19958,36 | 87,70 | 0,87 |
| Soltero/a | 3364 | 29,62% | 21406,38 | 18000 | 20867,97 | 97,48 | 0,49 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | | | | |

| **Cuadro 16. Distribución del ingreso de la ocupación principal según la Categoría ocupacional.** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Categoría\_ocup | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Cuenta propia | 2364 | 20,95% | 14044,99 | 10000 | 15086,03 | 107,41 | 0,80 |
| Obrero o empleado | 8583 | 76,07% | 30152,80 | 27000 | 23523,02 | 78,01 | 0,40 |
| Patrón | 336 | 2,98% | 36255,65 | 25000 | 51076,86 | 140,88 | 0,66 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | | | | |

| **Cuadro 17. Distribución del ingreso de la ocupación principal según la Intensidad de la ocupación.** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Intensidad | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Sobreocupado | 2319 | 20,42% | 32262,92 | 28000 | 25821,85 | 80,04 | 0,50 |
| Ocupado pleno | 5471 | 48,18% | 27983,49 | 25000 | 26053,23 | 93,10 | 0,34 |
| No\_trabajo | 2476 | 21,8% | 25480,10 | 23000 | 19642,93 | 77,09 | 0,38 |
| Subocupado | 1090 | 9,6% | 12093,22 | 8000 | 13335,47 | 110,27 | 0,92 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | | | | |

| **Cuadro 18. Distribución del ingreso de la ocupación principal según la Relación con el Jefe/a del hogar.** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Relación | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Jefe/a | 5773 | 50,84% | 30897,04 | 28000 | 27516,83 | 89,06 | 0,32 |
| Cónyuge | 2643 | 23,27% | 26977,53 | 24000 | 22832,30 | 84,63 | 0,39 |
| No\_fliar | 42 | 0,37% | 23083,33 | 17000 | 18249,60 | 79,06 | 1,00 |
| Madre/Padre | 45 | 0,4% | 22779,11 | 21600 | 16065,90 | 70,53 | 0,22 |
| Otros | 318 | 2,8% | 20411,01 | 18000 | 14935,16 | 73,17 | 0,48 |
| Hermano/a | 166 | 1,46% | 18356,02 | 15500 | 14680,53 | 79,98 | 0,58 |
| Hijo/a | 2259 | 19,89% | 18230,45 | 15000 | 15186,47 | 83,30 | 0,64 |
| Nieto/a | 110 | 0,97% | 16368,82 | 13450 | 14608,83 | 89,25 | 0,60 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | | | | |

| **Cuadro 19. Distribución del ingreso de la ocupación principal según el Tipo de establecimiento.** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Tipo | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Estatal | 3210 | 28,27% | 36229,29 | 30000 | 24951,87 | 68,87 | 0,75 |
| Otro | 127 | 1,12% | 28046,17 | 28000 | 17880,01 | 63,75 | 0,01 |
| Privado | 8019 | 70,61% | 22986,39 | 18000 | 23113,78 | 100,55 | 0,65 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | | | | |

A partir del Gráfico 4 y Cuadro 12 se observa que más de la mitad de las personas relevadas son hombres (55,64%) y que en promedio los mismos tienden a percibir un ingreso mayor por su ocupación principal que las mujeres ($29372,35 y $23543,33 respectivamente). A su vez, los ingresos de los hombres presentan una mayor variabilidad que el de las mujeres, aspecto que se registra al observar el CV de cada uno.

Con respecto al nivel educativo, aproximadamente el 43,49% de las personas cuentan con el secundario completo, mientras que un 28,73% con primaria completa y un 24,74% con universitaria completa. Por otro lado, se observa que a mayor nivel educativo las personas tienden, en promedio, a percibir ingresos más elevados (Gráfico 4 y Cuadro 13). Este aspecto también se observa al analizar cómo se distribuye el ingreso en función de la calificación del puesto de trabajo y de la intensidad de la ocupación. Aquellos puestos que requieren una mayor preparación tienden, en promedio, a percibir ingresos más altos. Los profesionales ganan un 62,17% más que los técnicos y estos últimos un 30,41% más que los operarios. A su vez, más de la mitad de las personas (52,04%) se desempeñan como operarios, mientras que sólo el 8,54% lo hace como profesionales (Gráfico 5 y Cuadro 14). En relación con la intensidad, se observa que casi la mitad de las personas (48,18%) son ocupados plenos, es decir que trabajaron más de 35 y menos de 45 horas semanales o bien menos de 35 horas pero que no desearon trabajar más, mientras que el 20,12% son sobreocupadas (trabajaron más de 45 horas semanales), el 9,6% subocupadas (trabajaron menos de 35 horas semanales, pero desearon trabajar más), y el restante 21,8% se trata de aquellos ocupados que no trabajaron en la semana. En relación con los ingresos, se observa que las ocupaciones de mayor intensidad (sobreocupado y ocupado pleno) son las que perciben ingresos más elevados (Gráfico 6 y Cuadro 17).

Por otra parte, el 76,07% de las personas relevadas se desempeña como obrera o empleada, mientras que aproximadamente el 3% lo hace como patrón y el 20,95% restante como cuentapropista. Estos últimos son los que en promedio perciben menores ingresos, seguidos por los empleados y luego por los patrones (Gráfico 6 y Cuadro 16). Además, el 70,61% se desempeña en el sector privado mientras que el 28,27% lo hace en el público, siendo en este último donde, en promedio, se perciben los ingresos más elevados (Gráfico 7 y Cuadro 19).

Por último, se observa que el 50,84% de las personas se perciben como jefe/as del hogar y, en promedio, son los que registran mayores ingresos, seguidos por los cónyuges quienes representan el 23,23% y ganan un 14,52% menos que los jefe/as (Gráfico 7 y Cuadro 18). Con respecto al estado conyugal, las personas casadas son las que perciben mayores ingresos, seguidas por aquellas unidas y solteras (Gráfico 5 y Cuadro 15).

La variable Categoría refiere al carácter ocupacional basado en el Clasificador Nacional de Ocupaciones (CNO), el mismo se agrega en grandes grupos ocupacionales que luego se desagregan en los grupos de carácter específicos[[1]](#footnote-1). Los niveles de la variable Categoría (51) del *dataset* refieren a esto último. En este caso, se procedió a generar una nueva variable que refiera a los grandes grupos ocupacionales (utilizando el diccionario del CNO) y evaluar cómo se distribuye el ingreso de la ocupación principal en función de cada grupo (Cuadro 20). A términos generales se observa que, aquellas personas que se desempeñan en ocupaciones de dirección, el 4,21% de las personas relevadas, son las que, en promedio, registran los mayores ingresos ($46550), seguidos por los que trabajan en ocupaciones auxiliares de producción de bienes y prestación de servicios (el 1,24% que reportan un ingreso medio de $38624,29), y por los que se desempeñan en la gestión administrativa, planificación y jurídico-legal (el 12,67% con un ingreso promedio de $37407,90).

| **Cuadro 20. Distribución del ingreso de la ocupación principal según el caracter ocupacional** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Categorías agrupadas | N | Freq | Media | Mediana | Desvío\_std | CV | Asimetría |
| Ocupaciones de dirección | 488 | 4,32% | 46550,00 | 35000 | 57014,21 | 122,48 | 0,61 |
| Ocupaciones auxiliares de la producción de bienes y la prestación de servicios | 140 | 1,24% | 38624,29 | 35000 | 23238,53 | 60,17 | 0,47 |
| Ocupaciones de gestión administrativa, de planificación, control de gestión y jurídico-legal | 1430 | 12,67% | 37407,90 | 32500 | 24550,36 | 65,63 | 0,60 |
| Ocupaciones de gestión presupuestaria, contable y financiera | 475 | 4,21% | 36280,45 | 33800 | 22508,08 | 62,04 | 0,33 |
| Ocupaciones servicios sociales básicos | 2116 | 18,74% | 33828,26 | 30000 | 19567,77 | 57,84 | 0,59 |
| Ocupaciones de la producción extractiva, energética, de construcción e infraestructura | 1037 | 9,19% | 22709,35 | 17000 | 23201,20 | 102,17 | 0,74 |
| Ocupaciones de comercialización,transporte, telecomunicaciones y almacenaje | 2312 | 20,48% | 22344,84 | 20000 | 18676,22 | 83,58 | 0,38 |
| Ocupaciones de la producción artesanal, industrial y de reparación de bienes de consumo | 1180 | 10,45% | 21367,88 | 18000 | 18708,42 | 87,55 | 0,54 |
| Ocupaciones agropecuarias, forestales, de la pesca y de la caza | 91 | 0,81 % | 18900,00 | 12900 | 14811,32 | 78,37 | 1,22 |
| Ocupaciones de servicios varios | 2021 | 17,9 % | 14369,51 | 12000 | 12046,77 | 83,84 | 0,59 |
| **Fuente: EPH 2º Trimestre 2020.** | | | | | | | |

1. **Modelización**
   1. *División en training y testing*

Para la definición del set de *training* y *testing* no se tomaron en cuenta las variables de “Nro\_hogar” y “Codusu” ya que se entiende que no son relevantes para predecir el ingreso. Por otro lado, se decidió incluir a la variable “Categoria\_agrup” que refiere a los grandes grupos ocupacionales, omitiendo a la de “Categoria” por una cuestión de interpretabilidad y dimensionalidad a la hora de modelizar. La primera de ellas cuenta con 11 niveles mientras que la segunda posee 51.

Se particionó el *dataset* en 70% para *training* y 30% para *testing*, optando por un diseño de remuestreo empleando *cross-validation* con 10 *folds* (k) y 3 repeticiones para el entrenamiento de los modelos[[2]](#footnote-2)*.* Esto último se realiza para conseguir estimaciones más certeras en el set de *training* antes de recurrir al conjunto de *testing.* La idea es ajustar y evaluar el modelo múltiples veces con distintos subconjuntos creados a partir de los datos de *training*, obteniendo en cada repetición una estimación del error. El promedio de todas las estimaciones tiende a converger en el valor real del error de *testing*.

* 1. *Preprocesado del dataset*

En primer lugar, se realizó una imputación de valores faltantes (que en el *dataset* responden a casos de no respuesta). Para la variable Ingreso se decidió imputarlos por la mediana de dicha variable, dada la asimetría que presenta la distribución y la existencia de valores atípicos.

En segundo lugar, se analizó si las variables presentaban varianzas de cero o próximas a cero, ya que, en el primer caso, no aportan información acerca de la variabilidad de la variable dependiente. Mientras que, en el segundo, en donde las variables toman pocos valores (con leve frecuencia), se corre el riesgo que puedan llegar a tener varianza cero cuando se dividen las observaciones por *cross-validation* o *boostrap*, por ejemplo.

A partir del Cuadro 21 se puede observar el análisis de la varianza de las variables del *dataset.* La función *nearZeroVar* del paquete *caret* identifica como variables potencialmente problemáticas aquellas que tienen un único valor (cero varianza) o que cumplen las dos siguientes condiciones:

* Ratio de frecuencias (freqRatio): ratio entre la frecuencia del valor más común y la frecuencia del segundo valor más común. Tiende a 1 si las frecuencias están equidistribuidas y a valores grandes cuando la frecuencia del valor mayoritario supera en demasía al resto. Por default: freqCut = 95/5.
* Porcentaje de valores únicos (percentUnique): número de valores únicos dividido entre el total de muestras (multiplicado por 100). Este porcentaje se aproxima a cero cuanto mayor es la variedad de valores. Valor por defecto: uniqueCut = 10.

Los valores lógicos indican que no hay variables con varianza cero o próxima a cero dentro del *dataset*.

|  |  | **Cuadro 21. Análisis de la varianza** | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Variable |  | freqRatio | percentUnique | zeroVar | nzv |
|  |  | Nivel\_ed |  | 1,606 | 0,033 | FALSE | FALSE |
|  |  | Relación |  | 2,172 | 0,066 | FALSE | FALSE |
|  |  | Sexo |  | 1,273 | 0,017 | FALSE | FALSE |
|  |  | Edad |  | 1,011 | 0,571 | FALSE | FALSE |
|  |  | Estado |  | 1,108 | 0,041 | FALSE | FALSE |
|  |  | Tipo |  | 2,403 | 0,033 | FALSE | FALSE |
|  |  | Categoría\_ocup |  | 3,947 | 0,025 | FALSE | FALSE |
|  |  | Intensidad |  | 2,207 | 0,033 | FALSE | FALSE |
|  |  | Total\_horas |  | 2,123 | 0,728 | FALSE | FALSE |
|  |  | Calificación |  | 2,594 | 0,058 | FALSE | FALSE |
|  |  | Ingreso |  | 1,732 | 2,597 | FALSE | FALSE |
|  |  | Categoría\_agrup |  | 1,057 | 0,091 | FALSE | FALSE |
| **Fuente: EPH 2º Trimestre 2020** | | | | | | | |

En tercer lugar, se aplicó el logaritmo en base 10 al Ingreso, dada la asimetría que presenta la distribución de la variable y la existencia de datos atípicos (extremos), para ello no se tuvieron en cuenta los casos que reportaron Ingreso = 0 (712 registros). Por otro lado, se decidió estandarizar las variables “Total\_horas” y “Edad” para que estén en las mismas unidades (escala Z), es decir que a cada variable se la dividió por su desviación típica después de haber sido centrada (restada por su media). Luego, se binarizaron las variables categóricas, se crearon nuevas variables *dummies* con cada uno de los niveles de estas. A este proceso también se le conoce como *one hot encoding[[3]](#footnote-3).*

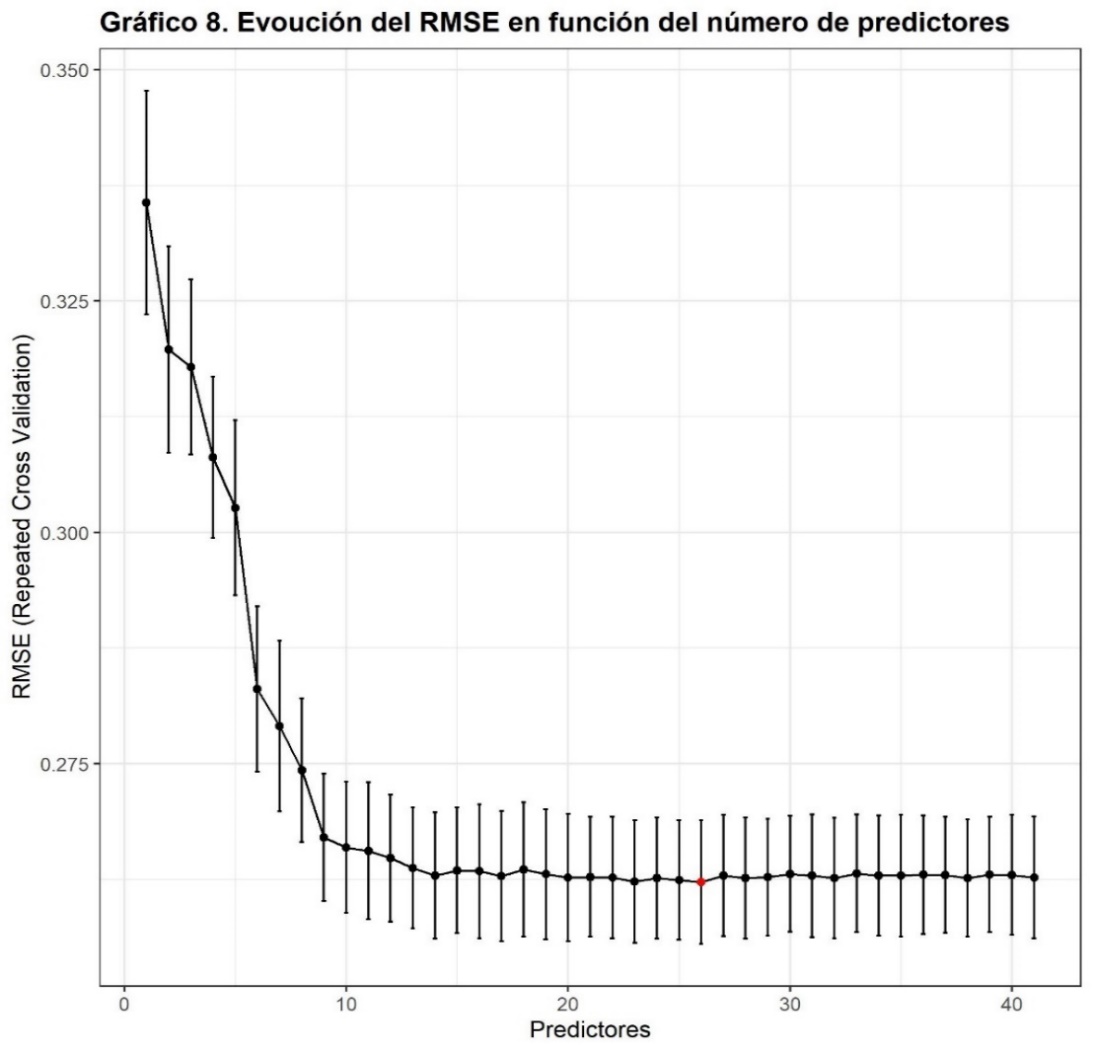
* 1. *Selección de predictores*

Existen diversos métodos, previo al ajuste del modelo, para reducir el número de predictores, los mismo pueden agruparse en: métodos *wrapper* y métodos de filtrado. El primero de ellos se basa en evaluar múltiples modelos, generados mediante la incorporación o eliminación de predictores, con la finalidad de identificar la combinación óptima que consigue maximizar la capacidad predictiva del modelo. El *Recursive Feature Elimination* (RFE) es un algoritmo englobado bajo los métodos *wrapper,* que a términos generales ajusta un modelo con todas las variables y luego va eliminando las más débiles (con menos importancia) hasta encontrar el número óptimo de las mismas. Este proceso lo realiza recursivamente, mediante *loops*, buscando eliminar las dependencias y colinealidades que puedan existir dentro del modelo. Para encontrar el número óptimo de variables, se utiliza *cross-validation* o *bootstrapping* junto a RFE para rankear diferentes subconjuntos de variables y seleccionar aquellos con mejor puntuación. El error final de cada modelo se estima agregando los errores obtenidos para cada conjunto de validación o *resampling* (de este modo se buscar reducir el posible *overfitting*).

Por otro lado, los métodos basados en el filtrado evalúan la relevancia de los variables fuera del modelo para, posteriormente, incluir únicamente aquellas que pasan un determinado criterio. Se analiza la relación que tiene cada predictor con la variable respuesta. Al igual que con los métodos *wrapper*, para evitar que la selección esté sumamente influenciada por los datos de *training* (*overfitting*), se itera el proceso mediante *cross-validation* o *bootstrapping.* En este caso para correr el RFE se estimó un modelo de *Random Fores*t (RF) y una estrategia de *Cross Validation* con 10 *folds* (K) y 2 repeticiones[[4]](#footnote-4).

A continuación, se presentan los resultados obtenidos al correr el RFE sobre el *dataset.* Se obtuvo que el mejor modelo es aquel que incorpora 26 predictores, siendo los 5 más importantes: “Categoria\_ocup\_Obrero o empleado”, “Total\_horas”, “Sexo\_Varon”, “Calificacion\_No calificados”, “Tipo\_Prviado” (Cuadro 22). El Gráfico 8 muestra cómo cambia el *Root Square Mean Error* (RMSE) cuando se van incorporando predictores al modelo, el punto rojo indica la cantidad de 26 predictores. Para cada conjunto de predictores el modelo de RF se estimó 10 veces (separando los datos en *training* y *testing*) y se repitió 2 veces, estimando el error en cada iteración y repetición para luego promediarlo y obtener el RMSE medio de cada conjunto.

| **Cuadro 22. 26 Predictores más influyentes estimados mediante RFE.** | |
| --- | --- |
| Predictores | Influencia |
| Categoria\_ocup\_Obrero.o.empleado | 69,081 |
| Total\_horas | 64,159 |
| Sexo\_Varon | 59,249 |
| Calificacion\_No.calificados | 38,565 |
| Tipo\_Privado | 38,527 |
| Edad | 37,763 |
| Categoria\_agrup\_ocupaciones.de.servicios.varios | 36,808 |
| Nivel\_ed\_Superior | 34,496 |
| Calificacion\_Profesionales | 32,198 |
| Categoria\_agrup\_ocupaciones.servicios.sociales.básicos | 28,105 |
| Calificacion\_Operativos | 27,748 |
| Nivel\_ed\_Secundaria | 27,679 |
| Intensidad\_Subocupado | 26,257 |
| Relacion\_Jefe.a | 25,567 |
| Calificacion\_Técnicos | 25,288 |
| Nivel\_ed\_Primaria | 25,230 |
| Relacion\_Hijo.a | 25,067 |
| Estado\_soltero.a | 24,396 |
| Categoria\_agrup\_ocupaciones.de.gestión.administrativa..de.planificación..control.de.gestión.y.jurídico.legal | 22,818 |
| Intensidad\_Ocupado.pleno | 22,746 |
| Categoria\_agrup\_ocupaciones.de.comercialización.transporte..telecomunicaciones.y.almacenaje | 19,829 |
| Intensidad\_Sobreocupado | 14,342 |
| Categoria\_agrup\_ocupaciones.de.la.producción.extractiva..energética..de.construcción.e.infraestructura | 11,911 |
| Categoria\_ocup\_Patron | 9,937 |
| Categoria\_agrup\_ocupaciones.de.dirección | 9,623 |
| Categoria\_agrup\_ocupaciones.de.gestión.presupuestaria..contable.y.financiera | 9,273 |
| **Fuente: EPH 2º Trimestre 2020.** | |

**

Si bien el mejor ajuste (menor RMSE) se logra con 26 predictores y basándose en el principio de parsimonia, según el cual, de entre un conjunto de modelos con similares capacidades predictivas, se prioriza aquel con menor complejidad, en este caso se mantendrán todos los predictores. Ya que, si sólo se eligen a los 26 con mayor influencia, se está seleccionando a ciertos niveles de las variables categóricas (eliminando el resto) y por tal, esto genera un cambio en el significado de las variables tal cual fueron relevadas. Por otra parte, el resultado de RFE sólo está mostrando que ciertos niveles de las variables categóricas contienen más información que otros, acerca de la variabilidad del ingreso de la ocupación principal. Dado que el fin último es la predicción, resulta pertinente dejar todos los predictores. Si se tratase de un análisis en donde prima la interpretabilidad, en ese caso, podría ser conveniente reducir la dimensionalidad del modelo para que sea más fácil su interpretación.

* 1. *Creación de modelos predictivos*

Se decidió estimar varios modelos para comparar sus errores de *training* y *testing* para luego seleccionar el de mejor performance, para el entrenamiento de dichos modelos se empleó una estrategia de *cross-validation* con 10 *folds* (k) y 3 repeticiones, tal cómo se especificó en la sección 2.1. De este modo, los modelos que se estimaron fueron: de la familia de las regresiones (Lineal Múltiple, *Ridge* y *Lasso*) y de los de árboles de decisión (CART, *Random Forest* y *Extreme Gradient Boost*). Para el caso de *Random Forest* y *Extreme Gradient Boost*, se estimaron sin tunear los hiperparámetros y luego tuneándolos para mejorar la bondad de ajuste de cada modelo. Antes de presentar los resultados obtenidos se desarrollan las características principales de cada uno de ellos[[5]](#footnote-5).

* + 1. *Regresión Lineal*

El modelo de regresión lineal se emplea para predecir variables del tipo cuantitativas, el mismo puede plantearse del siguiente modo:

Donde representa la variable dependiente (o de interés, es decir, aquella sobre la que se quiere analizar la variabilidad) y las variables independientes (regresoras o predictoras, es decir aquellas con las cuales se pretende explicar la variabilidad de ) registradas en la unidad experimental , y es el número de coeficientes de regresión incluido el intercepto (). Para poder predecir el valor medio de se deben conocer los parámetros poblacionales del modelo, como son desconocidos, se los estiman puntualmente a través de estadísticos que surgen de la muestra o del *training set* . Lo anterior se realiza a través de un método de estimación denominado *Mínimos Cuadrados Ordinarios* (MCO), que busca obtener los valores de los estadísticos que minimicen la suma de los cuadrados de las desviaciones (diferencias) entre los valores observados de y los predichos por el modelo de regresión (Suma de Cuadrados del Error, SCE). El criterio que se emplea se basa en la siguiente expresión:

Para que el modelo de regresión lineal sea válido se deben cumplir una serie de supuestos que se evalúan sobre los residuos (que son estimaciones del error del modelo):

* Homocedasticidad: Para cada valor de las variables independientes (o combinación de sus valores), la varianza es la misma.
* Independencia: los residuos no están correlacionados. El residuo para un determinado conjunto de valores de las variables independientes no está relacionado con el valor de residuo de ningún otro conjunto de valores.
* Normalidad: los residuos se distribuyen de forma Normal.
* Linealidad: el valor medio que se predice es modelado por una suma de parámetros.
  + 1. *Regresiones Ridge y Lasso*

Las regresiones *Ridge* y *Lasso (L2 y L1)* son métodos para regularizar imponiendo una penalización al tamaño de los coeficientes de la relación lineal entre las características predictivas y la variable dependiente. Dichos coeficientes, en *Ridge*, se estiman minimizando la siguiente expresión:

El primer término de la ecuación es igual al SCE de la regresión lineal, el segundo hace referencia a la penalización por la incorporación de coeficientes (que es equivalente al cuadrado de la magnitud de estos). es un parámetro que controla el grado de penalización, cuánto más grande más pequeños serán los coeficientes estimados (tienden a cero cuando ). Por otro lado, si se obtienen las mismas estimaciones que en un modelo de regresión lineal. se estima por fuera del modelo, por lo general mediante una estrategia de *cross-validation.*

Un aspecto que difiere de la regresión lineal es que los coeficientes estimados por *Ridge* o *Lasso* no son equivalentes de escala, es decir que al multiplicarlos por una constante su estimación puede diferir bastante en función de la unidad de medida. De este modo, la estimación del coeficiente *j-ésimo* *Ridge* no sólo dependerá del valor de sino también de la escala de medida del predictor, como también en algunos casos puede depender de la escala de los otros predictores. Por tal motivo, se recomienda aplicar *Ridge* o *Lasso* una vez estandarizadas las variables.

La principal diferencia entre *Ridge* y *Lasso* radica que en el segundo al incrementarseel valor de los coeficientes estimados pueden ser cero y, por tal, ser excluidos del modelo. En *Ridge* se reduce la magnitud de estos, pero nunca se hacen cero, es decir que, el modelo ajustado incorpora todos los coeficientes. Este aspecto no tiene tanta importancia cuando sólo interesa la predicción, pero si al momento de interpretar los modelos, especialmente cuando el *dataset* contiene muchas variables. De este modo, los coeficientes estimados en *Lasso* se obtienen minimizando la siguiente expresión:

* + 1. *CART*

CART es el término que se emplea para los algoritmos de árboles de decisión que pueden utilizarse en problemas de regresión como de clasificación. La representación se basa en un árbol binario, en el caso de uno de regresión, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal. Cuando se quiere predecir una nueva observación, se recorre el árbol acorde al valor de sus predictores hasta alcanzar uno de los nodos terminales. La predicción del árbol es la media de la variable respuesta de las observaciones de entrenamiento que están en ese mismo nodo terminal.

El entrenamiento de un árbol consta de dos etapas:

* División del espacio de los predictores (nodos terminales o regiones).
* Predicción de la variable dependiente en cada nodo terminal o región.

El criterio de división, en los árboles de regresión, se basa en encontrar aquellas regiones que, minimizan la sumatoria de las desviaciones al cuadrado entre las observaciones y la media de la región (Suma de Cuadrados de Residuos). Como no es posible considerar todas las particiones del espacio de los predictores, se recurre a lo que se conoce como *Recursive Binary Splitting.* Este método busca encontrar en cada iteración el predictor y el punto de corte tal que, si las observaciones se distribuyen en las regiones definidas, se minimiza la SCR. De este modo, el proceso sigue la siguiente secuencia:

1. Se inicia en lo más alto del árbol, donde todas las observaciones pertenecen a la misma región.
2. Se identifican todos los posibles puntos de corte para cada uno de los predictores. En el caso de predictores cualitativos, los posibles puntos de corte son cada uno de sus niveles. Para predictores continuos, se ordenan de menor a mayor sus valores, el punto intermedio entre cada par de valores se emplea como punto de corte.
3. Se calcula el SCR total que se consigue con cada posible división identificada en 2.
4. Se selecciona el predictor y el punto de corte que resulta en la menor SCR, es decir, que da lugar a las divisiones más homogéneas posibles. Si existen dos o más divisiones que consiguen la misma mejora, la elección entre ellas es aleatoria.
5. Se repiten de forma iterativa los pasos 1 a 4 para cada una de las regiones que se han creado en la iteración anterior hasta que se alcanza alguna norma de *stop*. Algunas de las más empleadas son: que ninguna región contenga un mínimo de *n* observaciones, que el árbol tenga un máximo de nodos terminales o que la incorporación del nodo reduzca el error en al menos un % mínimo.

El proceso de construcción de árboles tiende a reducir el error de *training*, es decir, el modelo se ajusta muy bien a las observaciones empleadas en el set de *training*, provocando, como consecuencia, *overfitting* al ser muy baja la performance al predecir sobre nuevas observaciones. Existen dos estrategias para prevenir dicho problema: limitar el tamaño del árbol (*early stopping*) o aplicar un método de podado (*pruning*). En la primera de ellas, se busca limitar la ramificación del árbol a través del cumplimiento o no de ciertas condiciones (observaciones mínimas por división y/o por nodo terminal, profundidad máxima, número máximo de nodos terminales, y reducción mínima en el error para que una división se realice). En la segunda, se parte de estructura de árboles grandes (sin condiciones) para luego podarlos y llegar a un árbol sub-óptimo que posea el menor error de *testing.* Para realizarlo se utiliza el método de penalización de *Complexity Pruning*, que resulta muy similar al empleado en *Ridge* y *Lasso,* donde se busca minimizar la SCR pero teniendo en cuenta la cantidad de nodos terminales que se van agregando.

* + 1. *Random Forest y Gradient Boosting*

Los modelos de *Random Forest* (RF) están formados por un conjunto (o *ensemble*) de árboles de decisión individuales, cada uno entrenado con una muestra aleatoria ligeramente distinta extraídas del set de *training* mediante [*bootstrapping*](https://en.wikipedia.org/wiki/Bootstrapping_(statistics))*[[6]](#footnote-6)*. En cada árbol, las observaciones se van distribuyendo por bifurcaciones (nodos) generando la estructura del árbol hasta alcanzar un nodo terminal. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo.

Los métodos de *ensemble* combinan múltiples modelos en uno nuevo con el objetivo de lograr un equilibrio entre el sesgo (*bias*) y la varianza. Por lo general, los árboles pequeños tienen poca varianza, es decir que al cambiar las observaciones de *training* el modelo no se modifica en demasía, pero si alto *bias* (no logran representar correctamente la relación entre los predictores y la variable respuesta). En contraposición, los árboles grandes tienden a ajustarse bien a los datos de *training*, teniendo poco sesgo, pero si mucha varianza.

Los tipos de *ensemble* más utilizados son:

* *Bagging:* Se ajustan múltiples modelos, cada uno con un subconjunto distinto del set de *training* (vía *bootstrapping*). Para predecir, se toma la media de todas las predicciones (variables continuas) o la clase más frecuente (variables categóricas). Los modelos RF usan esta técnica.
* *Boosting:* Se ajustan secuencialmente múltiples modelos sencillos, llamados *weak learners*, de forma que cada modelo aprende de los errores del anterior. Como valor final, se toma la media de todas las predicciones (variables continuas) o la clase más frecuente (variables cualitativas). *AdaBoost*, *Gradient Boosting* y *Stochastic Gradient Boosting* emplean esta técnica.

*Gradient Boosting* también es conocido como un algoritmo *Gradient descent*. Se trata de un procedimiento de optimización muy genérico. El objetivo consta en cambiar los diferentes hiperparámetros del modelo (utilizando el concepto de derivada) de forma iterativa con la intención de minimizar una función de coste. Mide el gradiente local de dicha función para un conjunto de hiperparámetros definidos, y luego va cambiando los valores de estos hasta encontrar el punto mínimo de la gradiente.

*Extreme Gradient Boosting* (XGB) emplea todos los principios de *Gradient Boosting* (este resulta una generalización de *AdaBoost)*, la principal diferencia radica en que XGB utiliza un modelo más formal para regularizar y así poder controlar el *overfitting*, proporcionando una mejor capacidad predictiva que el resto de los modelos basados en *boosting.*

* 1. *Ajuste de los modelos*

Antes de presentar los resultados acerca de los ajustes de cada modelo, se especifica, en los casos necesarios, el tuneo de parámetros e hiperparámetros que fue realizado. En primer lugar, para los modelos de *Ridge* y *Lasso* el fue estimado previo al ajuste del modelo mediante *cross-validation.* Para ello, se utilizó la función *cv.glmnet* que utiliza por default 10 *folds*.

Gráfico, Histograma

Descripción generada automáticamente

En el Gráfico 9 se observa cómo evoluciona el *Mean Square Error* (MSE) al variar el valor del . Las líneas punteadas indican, la de más a la izquierda, el valor del que minimiza el MSE, la otra es el valor de considerando un error standard. Los números de arriba refieren a la cantidad de predictores estimados que no son cero, al incrementarse el valor de se penaliza más y es por lo que, en *Lasso*, la cantidad de estos disminuye. En este caso, en *Lasso* sólo dos predictores se hacen cero cuando toma su mínimo (0,000226484): “Tipo Ns..Nr” y “Calificacion\_Operativos”.

Para el modelo de *Random Forest,* se brindó una grilla en donde se estimó el modelo utilizando los valores de 1 a 13 para *mtry,* y de5, 10, 15, 20, 25 para *min.node.size*. *Mtry* hace referencia al número de variables que se seleccionarán en cada partición de cada árbol del bosque. *Min.node.size* almínimo número de observaciones que debe tener cada nodo para que se produzca la división. Por defecto emplea 1 para problemas de clasificación y 10 para problemas de regresión. Por otro lado, el número de árboles se definió en 500 (el paquete *caret* utiliza este valor). En el Gráfico 10 se presentan los valores óptimos de los hiperparámetros: *mtry* = 11 y *min.node.size* = 25[[7]](#footnote-7).

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Para el caso del modelo *CART,* se brindó una grilla probando diferentes valores de *max.depth* (1, 2, 4, 8, 16). Este hiperparámetro controla la profundidad máxima del árbol creado. También se puede describir como la longitud del camino más largo desde la raíz del árbol hasta una hoja. Se considera que el nodo raíz tiene una profundidad de 0. El mejor ajuste se logró con un *max.depth* de 16 (Gráfico 11).

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Para el caso de *Extreme Gradiente Boost,* el grillado se definió del siguiente modo:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Hiperparámetros | | Valores | | Descripción |
| *Propuestos* | *Seleccionados* |
| *Nrounds:* | (1000; 1500) | | 1000 | Número máximo de iteraciones (modelos que forman el *ensemble*). |
| *Eta* | | (0,01; 0,05) | 0,01 | Controla el ratio de aprendizaje del modelo, es decir la tasa a la que el modelo aprende de los patrones del *dataset.* Valores pequeños demandan más recursos computacionales, por lo general se soporta con número de iteraciones mayores. Suele encontrarse entre 0,01 y 0,03 |
| *Max\_depth* | | (2; 4; 6) | 6 | Profundidad del árbol, mientras más profundo más complejo es y mayor probabilidad de *overfitting* |
| *colsample\_bytree* | | (0,5; 1) | 0,5 | Número de variables suministradas al árbol. Por lo general, está entre 0,5 y 0,9. |
| *Subsample* | | (0,5; 1) | 0,5 | Número de observaciones suministradas al árbol. |
| *Gamma* | | (0; 50) | 0 | Controla la regularización (*overfitting*). |
| *min\_child\_weight* | | (0; 20) | 20 | Mínimo número de instancias requeridas en un nodo hijo. |

* 1. *Comparación de modelos*

A partir del Gráfico 12 y 13 se puede observar la performance de los modelos ajustados, tanto en el set de *training* como de *testing*.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión, Gráfico de burbujas

Descripción generada automáticamente

En primer lugar, se destaca que el modelo de XGB es el que menor error de *testing* posee, seguido muy de cerca por el RF (con tuneo) y el RF (sin tuneo, es decir que se estimó con los valores de *default* de los hiperparámetros). Por otro lado, el modelo de CART es el que tuvo la peor performance (mayor media y mediana de RMSE) y, entre las regresiones, las diferencias resultan sumamente mínimas (Gráfico 12).

En segundo lugar, todos los modelos consiguen un mejor ajuste en el *set* de *training* que en el de *testing*, es decir que el error de *training* es menor que el de *testing* (lo cual resulta lógico). A su vez, las diferencias entre ambos errores son relativamente pequeñas (en especial en XGB y RF) lo que daría indicios de que no habría un problema de *overfitting*.

De este modo, el modelo de XGB es el que mejores resultados obtuvo, tanto en el conjunto de *testing* como en el de *training.* Dado que el fin último resulta en la predicción de los ingresos de la ocupación principal, se empleará dicho modelo para tal fin. Si se tratase de comprender e interpretar las relaciones entre las distintas variables independientes y la dependiente, en ese caso se optaría por modelos que resulten más sencillos para interpretar, como las regresiones. Otra forma de observar la bondad de ajuste de un modelo es a través de un diagrama de dispersión entre los valores observados y predichos por el modelo. En el Gráfico 15 se observa el ajuste de los modelos de XGB y RF luego de haber sido entrenados junto a una recta de 45 grados. Lo ideal es que los puntos estén lo más cercanos a la recta, esto indicaría que las diferencias entre el valor real observado y el predicho por el o los modelos es mínima. Por otro lado, se esperaría que la dispersión de los puntos por encima y por debajo de la recta sea lo más homogénea posible, mostrando que no habría problema de sub o sobreestimación. En este caso, en primer lugar, se aprecia que el ajuste de ambos modelos es muy similar, en segundo, que existe bastante variabilidad en torno al ajuste. En los extremos, valores bajos y altos de predichos, se aprecia una mayor variabilidad que en el segmento más del medio (entre 4 y casi 5 en los predichos). Lo anterior indica que el modelo es perfectible, tal vez se estén omitiendo variables que resultan significativas a la hora de predecir los ingresos de las personas (características de la vivienda, región o aglomerado de pertenencia, entre otras).

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

* 1. *Interpretación*

Tabla

Descripción generada automáticamente con confianza bajaPara poder interpretar el output que genera el modelo de XGB se parte de una concepción que va de lo general a lo más particular. A partir del Gráfico 16 es posible observar la contribución de cada predictor en el modelo. Dicha contribución es calculada a partir de la reducción que se genera en la Suma de Cuadrados de los Residuos al incorporar cada término al modelo (existen otras métricas para la estimación[[8]](#footnote-8)). Es así como se aprecia que, a términos generales, la cantidad de horas trabajadas en la semana anterior (en la ocupación principal), como si la persona es empleada, y la edad que posee resultan ser las características que mayor relevancia tendrían a la hora de explicar el logaritmo del ingreso de la ocupación principal de la persona.

Sin embargo, las medidas de importancia no dan una idea de la dirección promedio en que una variable afecta una función de respuesta. Simplemente establecen la magnitud de la relación de una variable con la respuesta en comparación con otras variables utilizadas en el modelo. Para poder entender cómo varía la variable respuesta en función de todas las variables se pueden utilizar las siguientes graficas: *Partial Dependece Plots* (PDP) y *Individual Conditional Expectation* (ICE). Estas técnicas muestran el cambio en el valor predicho a medida que las características específicas varían en su distribución marginal. En consecuencia, es posible comprender cómo cambia la variable respuesta a lo largo de la distribución de una variable en particular, pero esto sólo proporciona una comprensión global de estas relaciones en los datos observados.

Para aplicar dichas técnicas, junto con LIME (que se verá a continuación), es recomendable colapsar las variables *dummies* en categóricas. A su vez, se transforman las variables independientes numéricas (que estaban estandarizadas) a su escala original para que sea más fácil la interpretación[[9]](#footnote-9).

El Gráfico 17 muestra el PDP para las dos variables numéricas (“Total\_horas” y “Edad”) y para dos categóricas (“Nivel\_ed” y “Sexo”). En el primer caso, se observa que a medida que aumentan las horas trabajadas, el logaritmo del ingreso que reciben las personas (según el modelo) se incrementa hasta un cierto punto, para luego estabilizarse (a partir de las 70 horas aproximadamente el ingreso no se modifica). En el segundo caso, ocurre algo similar, pero en vez de estabilizarse el logaritmo del ingreso a partir de un cierto umbral, el mismo comienza a caer afianzándose en un menor nivel. (se observa que el máximo de logaritmo del ingreso se percibe, en promedio, a los casi 60 años).

Dichos comportamientos pareciesen ser bastante lógicos, en tanto se esperaría que, a mayor cantidad de horas trabajadas, en promedio el ingreso aumente, pero este incremento no sería lineal en todo momento, sino más bien tendría una dinámica asintótica luego de cierto nivel. Por otro lado, a medida que una persona crece (capacitándose y acumulando experiencia laboral) es lógico suponer que su ingreso aumente, pero este incremento no es lineal, sino que, a partir de cierto momento, el aumento es cada vez menor.

Con respecto a las variables de Nivel Educativo y Sexo los resultados son los esperables, a mayor nivel educativo mayor el logaritmo del ingreso que se percibe, y los varones, en promedio, ganan más que las mujeres.

Gráfico

Descripción generada automáticamente

Los PDP son un método global, no se enfocan en instancias específicas, en tanto sólo miran el promedio general. Los ICE permiten visualizar la dependencia de la predicción en una característica para cada instancia por separado, lo que da como resultado una línea por instancia, en comparación con una línea general en los PDP.

Gráfico

Descripción generada automáticamente

Al graficar el ICE (Gráfico 18), para las dos variables numéricas, se aprecia que se mantiene la tendencia general descrita en el PDP. La dinámica que presenta cada línea (sería cada persona) tiende asemejarse bastante a la agregada (representada en PDP), aunque en ciertos casos se observan algunas diferencias, las mismas son mínimas.

A su vez, es posible conocer la influencia de cada predictor por sobre la variable respuesta en cada registro (observación) a través de una técnica llamada *Local Interpretable Model-Agnostic Explanations[[10]](#footnote-10)* (LIME). La misma es una técnica de visualización que ayuda a explicar las predicciones individuales. Al ser independiente del modelo (de ahí la referencia de *Agnostic*), puede aplicarse a cualquier modelo de clasificación o regresión supervisada. Según los creadores de LIME, cualquier modelo complejo es lineal a una escala local y, por tal, es posible ajustar un modelo simple en torno a una sola observación que imitará cómo se comportará el modelo global en esa localidad. Luego, el modelo simple se puede usar para explicar localmente las predicciones del modelo más complejo.

De forma esquemática, los pasos seguidos por LIME son:

1. Dada una observación a explicar, se permuta mediante ligeras modificaciones en los valores de sus predictores para obtener observaciones nuevas.
2. El modelo complejo predice las observaciones permutadas.
3. Calcula la distancia entre las observaciones permutadas y la original y las convierte a medida de similitud.
4. Selecciona los predictores que mejor describan las predicciones anteriores.
5. Ajustar un modelo sustituto interpretable a las observaciones permutadas que explique el resultado del modelo global usando los predictores seleccionados.
6. Extraer los pesos de los predictores del modelo simple para para explicar el comportamiento local del modelo global (complejo).

A continuación, se presentan los resultados obtenidos para las primeras tres observaciones del *set* de *testing.*

Gráfico

Descripción generada automáticamente

Para cada observación se reporta el valor predicho (en este caso el logaritmo del ingreso) y la bondad de ajuste (*Explanation Fit*) del modelo simple lineal ajustado a la localidad de la observación original. A su vez, se muestran los cinco predictores más importantes que intervinieron en la predicción, en azul aquellos que correlacionaron positivamente con la variable dependiente y en rojo los que lo hicieron negativamente.

De este modo, para la primera observación, que responde a una mujer que trabajó más de 40 horas en la semana anterior (en su ocupación principal), que es empleada en un establecimiento estatal y que cuenta con estudios de nivel superior (entre otras características), el modelo predice un logaritmo del ingreso de la ocupación principal de 4,5661 (es decir de aproximadamente $35164,19). De dichas características, las de mayor influencia en la predicción fueron: haber trabajado más de 40 horas y ser empleada. Por lo contrario, el ser mujer afectó negativamente en la predicción.

En el caso de la segunda observación, que la persona sea cuentapropista, trabajé en un establecimiento privado y se desempeñe en un puesto de trabajo operativo, son características que afectaron negativamente, mientras que sea varón lo hizo positivamente. A diferencia de la observación 1, en la 2 la bondad del ajuste es menor, por tal el resultado de la interpretación puede ser menos confiable.

Por último, para la tercera observación se aprecia que el ser empleado, trabajar en un establecimiento estatal, ser varón y tener una edad en el rango de 41-51 años, fueron factores que incidieron positivamente en la predicción, mientras que el ser hijo/a lo hizo negativamente.

1. **Referencias Bibliográficas**

* INDEC (2018). *Clasificador Nacional de Ocupaciones (CNO)*. Buenos Aires.
* INDEC (2020). *Encuesta Permanente de Hogares. Diseño de registro y estructura para las bases preliminares Hogar y Personas*. Buenos Aires.
* Faraway. J. (2006). *Linear Models with R.* New York: A CRC Press Company.
* Garibaldi, et al. (2019). *Modelos estadísticos en lenguaje R*. Editorial UNRN.
* Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* *7th ed*. New York: Springer.
* Riberio, M., Samer, S. y Guestrin, C. (2016). *“Why Should I Trust You?”: Explaining the Predictions of Any Classifie*r. Disponible en: https://arxiv.org/abs/1602.04938v3 (consultado el 02/03/22).

1. Para más información consultar en: https://www.indec.gob.ar/ftp/cuadros/menusuperior/clasificadores/definiciones\_conceptuales\_cno.pdf [↑](#footnote-ref-1)
2. *Cross Validation* con 10 folds y 3 repeticiones, implica ajustar y evaluar el modelo 30 veces, cada vez con una partición distinta, más un último ajuste con todos los datos de entrenamiento para crear el modelo final. [↑](#footnote-ref-2)
3. Como ejercicio, en una primera instancia, se decidió incorporar la variable ya que se suele utilizar a la hora de predecir el ingreso de las personas (generalmente en regresiones). Pero al comparar los ajustes de los modelos planteados, la incorporación de dicha variable no generó una mejora considerable en la performance de estos, salvo en las regresiones (lineal, *Lasso* y *Ridge*), pero no así en los modelos de *ensemble* (*Random Forest* y *Xgboost*) los cuales tuvieron mejor performance. Por tal motivo, no se la incluyó en la modelización final. [↑](#footnote-ref-3)
4. Dado que el tiempo de ejecución fue de 7 hs (empleando un procesador de i7 16 gb RAM y 500 gb SSD) se adjunta el modelo rf\_rfe en formato RDS. [↑](#footnote-ref-4)
5. La sección se desarrolló en base a: Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* 7th ed. New York: Springer y Faraway. J. (2006). *Linear Models with R*. New York: A CRC Press Company y Garibaldi, et al. (2019). *Modelos estadísticos en lenguaje R*. Editorial UNRN. [↑](#footnote-ref-5)
6. Método de remuestreo (aleatorio y con reposición) empleado para aproximar la distribución de un estadístico. La idea básica de *bootstrap* es que la inferencia estadística puede modelarse mediante un nuevo muestreo de los datos de la muestra y realizando la inferencia sobre una muestra (sería la “población”) a partir de datos remuestreados. [↑](#footnote-ref-6)
7. Se adjunta el modelo (“mod\_rf\_tun”) en formato RMD. [↑](#footnote-ref-7)
8. Al respecto, https://www.rdocumentation.org/packages/caret/versions/6.0-90/topics/varImp. [↑](#footnote-ref-8)
9. Estas modificaciones no generan cambios en el ajuste del modelo, tampoco en la optimización de los hiperparámetros. [↑](#footnote-ref-9)
10. El desarrollo de dicha técnica puede encontrarse en: Riberio, M., Samer, S. y Guestrin, C. (2016). “*Why Should I Trust You?”: Explaining the Predictions of Any Classifier*. Disponible en: <https://arxiv.org/abs/1602.04938v3> (consultado el 02/03/22). [↑](#footnote-ref-10)