Trabajo Fin de Máster

Estudio estadístico sobre el mercado laboral en España: análisis de la desigualdad entre hombres y mujeres.

Carolyn Jafreissy Bautista Del Orbe

Máster en Técnicas Estadísticas

Curso 2019-2020

Propuesta de Trabajo Fin de Máster

Título en galego: Estudo estatístico do mercado de traballo en España: análise da desigualdade entre homes e mulleres.
Título en español: Estudio estadístico sobre el mercado laboral en España: análisis de la desigualdad entre hombres y mujeres.
English title: Statistical study on the labor market in Spain: analysis of inequality between men and women.
Modalidad: Modalidad A
Autor/a: Carolyn Jafreissy Bautista Del Orbe, Universidad de La Coruña
Director/a: Antonio Vaamonde Liste, Universidad de Vigo
Breve resumen del trabajo: En este trabajo se utilizan microdatos de diferentes encuestas públicas (Encuesta de Población Activa, Encuesta de Estructura Salarial) para el estudio del mercado laboral español, con especial énfasis en el análisis de la desigualdad entre hombres y mujeres, mediante la aplicación de distintas técnicas estadísticas de análisis multivariante y minería de datos.

Don Antonio Vaamonde Liste, profesor de la Universidad de Vigo, informa que el Trabajo Fin de Máster titulado

Estudio estadístico sobre el mercado laboral en España: análisis de la desigualdad entre hombres y mujeres.

fue realizado bajo su dirección por doña Carolyn Jafreissy Bautista Del Orbe para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En A Coruña, a 6 de Septiembre de 2020.

El director:

Don Antonio Vaamonde Liste

La autora:

Doña Carolyn Jafreissy Bautista Del Orbe

Agradecimientos

En primer lugar, a Dios por obrar de manera especial en mi vida, por estar conmigo en todo momento y ser mi guía cuando no sabía que camino escoger, por sus infinitas bendiciones y hacer que este sueño hoy sea una realidad.

A mi familia que aun en la distancia siempre me han brindado su apoyo, en especial a mi madre Gladys Del Orbe, por haberme instruido por el camino correcto, por motivarme a continuar cuando sentía desvanecer, por apoyarme en todos mis proyectos y decisiones, por ser mi inspiración y mi mejor ejemplo de que todo en la vida es posible de lograr independientemente de las adversidades que se puedan presentar.

A mis amigas y compañeras del máster (Jomayra y Lady) por su apoyo incondicional durante todo el curso y las buenas experiencias vividas.

En especial a mi director Antonio Vaamonde Liste, por su tiempo y dedicación, siempre dispuesto a guiarme, corregirme y revisar el trabajo en todo momento que lo necesitaba estuvo ahí, su ayuda ha sido muy importante para la realización de este trabajo de fin de máster.

Gracias a todas las personas que me brindaron su apoyo y creyeron en este sueño.

Índice general

Resumen

Resumen en español

El presente trabajo tiene por objetivo analizar el mercado laboral español, haciendo especial énfasis en la desigualdad entre hombres y mujeres. En primer lugar, se tratará de identificar las variables más importantes que nos permitan explicar la situación en que se encuentran los encuestados en el mercado laboral y de esta forma conocer cuáles son los factores de riesgo para la situación de parado. Por otro lado, interesa conocer y cuantificar los factores que influyen en el salario y la brecha salarial entre hombres y mujeres.

Para ello se disponen de datos de la Encuesta de Población Activa (EPA) del tercer trimestre de 2019 (2019-T3) del mercado español. Además, se utilizan los datos de la Encuesta de Estructura Salarial (EES) española, especificamente los microdatos de la Encuesta cuatrienal de Estructura Salarial del año 2014. Ambas se descargan desde la página del Instituto Nacional de Estadística (INE), siendo las más actuales disponibles al momento de la búsqueda.

En ambos casos se aplicarán herramientas de análisis exploratorio y descriptivo de datos e inferencia, aplicando modelos de regresión logística, regresión lineal multivariante y técnicas propias de datos multivariantes. Al final se aplican técnicas de minería de datos con el objetivo de clasificar los individuos en parados y ocupados.

English abstract

The objective of this work is to analyze the labor market in Spain, with special emphasis on inequality between men and women. In the first place, an attempt will be made to identify the most important variables that make it possible to explain the current situation in which the respondents find themselves in the labor market and thus to know what the risk factors for unemployment are. On the other hand, it is interesting to know and quantify the factors that influence wages and the wage gap between men and women.

For this, data are available from the following surveys: Encuesta de Población Activa (EPA) for the third quarter of 2019 (2019-Q3) and Encuesta de Estructura Salarial (EES), specifically the microdata from the 2014 four-year. Both are downloaded from the National Institute of Statistics (INE) website, being the most updated available at the time of the search.

In both cases, tools of exploratory and descriptive analysis of data and inference are applied, as well as logistic regression models, multivariate linear regression and techniques of multivariate data. In the end, data mining techniques are applied to classify people as unemployed and employed.

Capítulo 1

Introducción

El trabajo es una de las facetas fundamentales de nuestra vida a la que dedicamos gran esfuerzo y que ocupa una parte importante de ella. En general, la vida laboral de una persona empieza con la finalización de los estudios o formación inicial, entre los 16 y 24 años, y se extiende hasta la jubilación, entre los 60 y los 65 años.

De acuerdo al SEPE (2019), en su informe del mercado de trabajo estatal, en el año 2018 las principales variables del mercado de trabajo tuvieron un comportamiento positivo. La tasa de paro descendió por sexto año consecutivo hasta situarse en el 14.45 %. La tasa de empleo se incrementó por quinto año situándose en el 50.14 %. Mientras que el dato menos positivo es el relacionado con la tasa de actividad que lleva descendiendo desde el año 2011 y se situó en el cuarto trimestre de 2018 en el 58.61 %.

La tasa de paro presentó fuertes incrementos entre los años 2007 y 2012 de 17.2 puntos porcentuales (p.p.). Las continuas bajadas de la tasa de paro de los últimos años solo han permitido compensar una parte de los fuertes incrementos de años anteriores.

A pesar de que en las últimas décadas se han ido produciendo grandes cambios en la sociedad en los que la mujer ha sido protagonista, y a pesar de que la participación de éstas en el mercado de trabajo ha ido aumentando, según el SEPE (2019), las tasas de actividad, empleo y paro presentan diferencias significativas entre hombres y mujeres. Las tasas de actividad son superiores en los hombres. Esta diferencia oscila, en los últimos cinco años, de los 11.37 p.p. en 2018 a los 12.05 p.p. de 2014. Esas relevantes diferencias en las tasas de actividad de las mujeres y los hombres tienen una gran estabilidad.

Las tasas de empleo son más de 10 p.p. superiores en los hombres. Para el cuarto trimestre de 2018 la diferencia fue de 11.7 p.p. La tasa de paro de las mujeres fue en los últimos cinco años más elevada que la de los hombres. Las diferencias en las tasas de paro de hombres y mujeres oscilaron entre los 1.94 p.p. del año 2014 y los 3.39 p.p. del año 2018.

En 2016 el salario medio anual de los hombres fue de 25,924€ y el de las mujeres de 20,131€. La brecha salarial de género[1], medida en términos de salario medio anual, fue del 22.35 %.

La reducción de la conocida "brecha salarial" es uno de los temas de más actualidad y uno de los principales retos a los que se enfrenta el mercado laboral tanto a nivel nacional como internacional. Según el CEOE (2019) durante las dos últimas décadas en España se ha avanzado de forma positiva en materia de igualdad de género en el mercado laboral, aumentando la tasa de participación de la mujer y reduciendo las diferencias salariales existentes entre hombres y mujeres, aunque, al igual que en el resto de países europeos, aún queda trabajo por hacer para fomentar la igualdad real entre ambos colectivos en la sociedad, en general, y en el mercado laboral, en particular.

En este trabajo se pretende estudiar la población económicamente activa del mercado laboral español. Específicamente interesa analizar las variables sociodemográficas más relevantes de los individuos, con el objetivo

[1] La Comisión Europea define la brecha salarial de género como "la diferencia relativa en el ingreso bruto promedio de mujeres y hombres dentro de la economía en su conjunto".

de explicar la relación de estas características con su situación actual en el mercado laboral, haciendo especial énfasis en el sexo. Interesa, además, estudiar los factores que influyen en la diferencia salarial entre hombres y mujeres. Para ello se utilizarán la Encuesta de Población Activa correspondiente al tercer trimestre de 2019 y la Encuesta de Estructura Salarial del año 2014.

Para llevar a cabo esta tarea, en el Capítulo 2 se desarrollan algunas consideraciones básicas a tener en cuenta sobre las encuestas utilizadas y se describen las variables a estudiar. En el Capítulo 3 se introducen las técnicas estadísticas que se van a emplear para el análisis de los datos. En los Capítulos 4 y 5 se realiza un análisis exploratorio y se aplican las técnicas explicadas en el Capítulo 3. A continuación, se utilizan técnicas de minería de datos para clasificar los individuos en parados y ocupados. Por último, en el Capítulo 7 se detallan las conclusiones finales.

Capítulo 2

Descripción y procesamiento de datos

En este trabajo se analizan dos bases de datos diferentes. El primer caso se corresponde con el análisis de la Encuesta de Población Activa del tercer trimestre de 2019[1]. En el segundo caso se analiza la Encuesta cuatrienal de Estructura Salarial de 2014[2].

En este capítulo se realiza una breve descripción de las encuestas y las variables utilizadas en el estudio.

2.1. Datos EPA (2019-T3)

La EPA se inició en 1964. Se trata de una investigación por muestreo de periodicidad trimestral, dirigida a la población que reside en viviendas familiares del territorio nacional y cuya finalidad es averiguar las características de dicha población en relación con el mercado de trabajo.

La finalidad principal de la encuesta es conocer la actividad económica en lo relativo a su componente humano. Está orientada a dar datos de la fuerza de trabajo y de sus diversas categorías (ocupados, parados), así como de la población ajena al mercado laboral (inactivos) y a obtener clasificaciones de estas categorías según diversas características. También posibilita confeccionar series temporales homogéneas de resultados. Además, al ser las definiciones y criterios utilizados coherentes con los establecidos por los organismos internacionales que se ocupan de temas laborales, permite la comparación con datos de otros países.

Hay que tener en cuenta una serie de definiciones[3] que ayudan a clasificar a cada uno de los encuestados. De tal forma que se considera:

Población económicamente activa: Conjunto de personas de 16 o más años que durante la semana de referencia suministran mano de obra para la producción de bienes y servicios económicos o que están disponibles y hacen gestiones para incorporarse a dicha producción.

Se divide en:

- ***Ocupados***: son todas aquellas personas de 16 años o más que, según los criterios de la Organización Internacional del Trabajo (OIT), durante la semana de referencia tuvieron un empleo por cuenta ajena, asalariado, o ejercieron una actividad por cuenta propia, trabajadores por cuenta propia.

  Se clasifica a la persona como ocupada cuando en la semana de referencia ha trabajado al menos una hora a cambio de una remuneración, salario, beneficio empresarial o ganancia familiar, o ha estado ausente del trabajo pero mantiene un fuerte vínculo con dicho empleo.

[1] Los microdatos de la encuesta se pueden encontrar en el siguiente enlace: http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=resultados&secc=1254736030639&idp=1254735976595

[2] Los microdatos de la encuesta se pueden encontrar en el siguiente enlace: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177025&menu=resultados&idp=1254735976596

[3] Las definiciones están basadas en las recomendaciones aprobadas por la Organización Internacional del Trabajo (OIT) en la Decimotercera y Decimosexta Conferencia Internacional de Estadísticos del Trabajo (Ginebra, 1982 y 1998, respectivamente).

- ***Parados***: son todas aquellas personas de 16 o más años que durante la semana de referencia se encuentren sin trabajo, hayan buscado activamente trabajo y estén disponibles para trabajar.

  Además, se considerarán parados aquéllos que han estado sin trabajo en la semana de referencia pero no buscan empleo porque han encontrado uno al que se incorporarán dentro de los tres meses posteriores a la semana de referencia.

Población inactiva: Comprende a todas las personas de 16 o más años no clasificadas como ocupadas ni paradas durante la semana de referencia.

Comprenden a personas que no trabajan, aquellas que aunque están en disposición de trabajar no buscan trabajo activamente o aquellos que sí buscan trabajo pero no estaban disponibles de forma inmediata. Por ejemplo, jubilados, personas dedicadas al hogar, estudiantes.

2.1.1. Descripción de variables

Los datos originales contienen un total de 163,365 observaciones y 93 variables.

```
dim(data_2019T3)
```

```
[1] 163365     93
```

La variable **`AOI`**, que será el objeto de estudio, contiene la clasificación de los entrevistados por relación con la actividad económica.

En esta variable clasifican a la población mayor o igual a 16 años de la siguiente manera:

Tabla 2.1. Clasificación entrevistados por act. económica, según criterios OIT.

Cód.	Descripción
03	Ocupados subempleados por insuficiencia de horas
04	Resto de ocupados
05	Parados que buscan primer empleo
06	Parados que han trabajado antes
07	Inactivos 1 (desanimados)
08	Inactivos 2 (junto con los desanimados forman los activos potenciales)
09	Inactivos 3 (resto de inactivos)

En este trabajo nos concentraremos solo en analizar la población activa, por lo que se crea una nueva variable, **`Activos`**, con las categorías "ocupados" y "parados", trabajando con los niveles 03, 04 y 05, 06 de forma conjunta. Si en algún caso nos interesara analizar la variable de forma desagregada utilizaremos la variable original **`AOI`**.

Con el objetivo de analizar la situación del mercado laboral de España, en el tercer trimestre de 2019, se explicará la situación en que se encuentre cada individuo con respecto a variables como:

- Personales y demográficas: **`Sexo, Grupos de edad, Estado civil, Comunidad autónoma`**.
- Laborales y formación: **`Nivel de estudios, Ocupación, Tipo de contrato, Tipo de jornada`**.

Luego de filtrar la población mayor o igual a 16 años, seleccionar los ocupados y parados, convertir las variables a estudiar en factores y etiquetar cada una de sus categorías, aplicamos la función `str()` a los

datos. Esta función muestra la estructura interna de un objeto y nos describe el tipo de variables en el *data frame*.

Los datos a utilizar se reducen a 74,862 observaciones y 12 variables.

```
str(data_activos)
```

```
'data.frame': 74862 obs. of  12 variables:
 $ ACTIVOS  : Factor w/ 2 levels "ocupados","parados": 1 1 1 1 1 1 1 1 1 1 ...
 $ SEXO     : Factor w/ 2 levels "hombre","mujer": 1 2 1 2 1 2 1 2 2 1 ...
 $ EDAD     : Factor w/ 11 levels "16 a 19","20 a 24",..: 5 7 7 6 7 5 8 8 4 5 ...
 $ ECIVIL   : Factor w/ 4 levels "Casado","Sep/divorciado",..: 3 2 1 1 3 3 3 3 3 1 ...
 $ CCAA     : Factor w/ 19 levels "Andalucía","Aragón",..: 19 19 19 19 19 19 19 19 19 19 ...
 $ ESTU     : Factor w/ 8 levels "  ","AN","P1",..: 6 8 8 8 8 8 8 8 8 8 ...
 $ NAC      : Factor w/ 3 levels "Doble nac.","Española",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ OCUP     : Factor w/ 10 levels "Directores y gerentes",..: 4 9 5 5 1 6 2 6 6 10 ...
 $ SITU_PROF: Factor w/ 7 levels "Asalariado sector privado",..: 1 1 2 1 1 7 2 1 1 1 ...
 $ TIPOCON  : Factor w/ 2 levels "Indefinido","Temporal": 1 1 1 1 1 NA 2 1 1 1 ...
 $ TIPOJOR  : Factor w/ 2 levels "Completa","Parcial": 1 2 1 1 1 1 1 1 1 1 ...
 $ MOT_JP   : Factor w/ 8 levels " No quiere jornada completa ",..: NA 6 NA NA NA NA NA NA NA NA ...
```

2.2. Datos EES (2014)

La Encuesta cuatrienal de Estructura Salarial se realiza en el INE desde el año 2002 con criterios de metodología y contenidos comunes en el marco de la Unión Europea (UE), con el fin de obtener unos resultados comparables sobre el nivel, la estructura y distribución del salario entre sus Estados Miembros, según lo previsto en el Reglamento del Consejo 530/1999.

La novedad principal que aporta frente a las otras encuestas tradicionales de coste laboral es que, siendo igualmente la unidad informante la empresa, se recogen los salarios en un cuestionario de forma individualizada para cada trabajador. Se dirige a todos los trabajadores que han estado de alta en la Seguridad Social durante todo el mes de octubre del año de referencia.

Junto a los salarios se incluye una gran cantidad de variables relacionadas con el trabajador y con el puesto de trabajo que ocupa. Gracias a esto, es posible establecer relaciones entre el salario y algunas variables que pueden contribuir a determinar su cuantía, como el nivel de estudios alcanzado, la antigüedad, el tipo de contrato o la ocupación, entre otras. Por tanto, la encuesta no sólo proporciona valores de salarios medios sino también de la distribución estadística de los mismos, y en consecuencia, una medida de su desigualdad.

Dentro de las percepciones salariales destacan los siguientes conceptos:

Salario base: parte fundamental y fija del salario que se define como el mínimo de retribución acordado en los convenios colectivos y calculado generalmente en euros/mes o euros/día. Cuando no hay convenio colectivo ni otro acuerdo entre empleador y empleado se entiende que éste es el Salario Mínimo Interprofesional (SMI).

Pagos por horas extraordinarias: corresponde a los pagos por horas extras tanto estructurales como no estructurales. Remuneran por tanto el mayor esfuerzo que representa el trabajo adicional realizado fuera de la jornada habitual.

Complementos salariales: conjunto de retribuciones pagadas por encima de las retribuciones básicas (salario base y pagas extraordinarias) que el empleador abona habitualmente previo pacto en convenio colectivo. Entre estos destacan: los complementos personales, donde se valora la antigüedad o vinculación continuada del trabajador a la empresa, la cualificación individual mediante la titulación académica o profesional, los conocimientos de idiomas, informática, etc.; los complementos relacionados con el puesto de trabajo, que se concretan en las específicas circunstancias en que se desarrolla el trabajo. Existen también los pluses de

nocturnidad, los de trabajo en días festivos, así como los pluses de peligrosidad, penosidad o toxicidad. Otros pluses son los complementos por calidad y cantidad de trabajo, que remuneran la asistencia y puntualidad, y los incentivos de productividad que premian el rendimiento laboral por encima de unos mínimos pero que se perciben todos los meses.

Pagos extraordinarios: recogen todos los pagos de vencimiento superior al período corriente de pago (el mes por regla general) aunque sean de carácter irregular. Se distinguen tres categorías fundamentales:

- ***Pagos extraordinarios fijos:*** fundamentalmente lo forman las gratificaciones extraordinarias o pagas de Navidad y de verano (reguladas por ley) y las gratificaciones por beneficios.
- ***Pagos extraordinarios variables:*** pagos por incentivos o resultados, son pagos que están ligados a resultados individuales o de la empresa.
- ***Pagos en especie (a partir de 2006):*** se trata de la valoración de todas las percepciones salariales que se entregan al trabajador usando cualquier medio distinto del dinero, como bienes, derechos, beneficios o servicios como contraprestación de los servicios laborales.

2.2.1. Descripción de variables

La variable a explicar es la **Ganancia media por hora**, necesaria para poder estudiar en condiciones de igualdad a los trabajadores con distintas jornadas, especialmente al trabajador a tiempo parcial.

Se calcula a partir de la información de la EES con la siguiente formulación, según la metodología utilizada por el INE:

Ganancia media por hora= Ganancia mensual / Horas trabajadas del mes de referencia (normales y extraordinarias).

Donde:

Ganancia mensual = Salario bruto mensual del mes de Octubre = Salario Base del mes de Octubre + Complementos salariales del mes de Octubre + Pagos por horas extraordinarias del mes de Octubre + Paga extraordinaria del mes de Octubre

Horas trabajadas mensuales = Horas normales de trabajo en el mes de octubre + Horas extraordinarias realizadas en el mes de octubre

Horas normales de trabajo en el mes de octubre = Jornada semanal pactada (horas + minutos/60)*4.35.

Para explicar el comportamiento de las Ganancias por horas utilizamos las variables que hacen referencia a distintas características personales y laborales de los trabajadores que se quieren estudiar. En particular:

- Personales y formación: **Sexo, Edad, Nivel educativo**.
- Laborales: **Antigüedad en la empresa, Tipo de jornada laboral, Duración del contrato, Responsabilidades organizativas y/o de supervisión y Ocupación.**

La mayoría de las variables explicativas incluidas en los modelos son variables discretas. Sin embargo, la antigüedad es continua y para introducir esta variable al análisis también se categoriza con el objetivo de analizar y estimar la brecha salarial por el sexo para las diferentes categorías que puedan ser de interés.

```
str(data_ees)
```

```
'data.frame': 209436 obs. of  9 variables:
 $ SAL      : num  38.7 20.3 15.2 18.3 33.5 ...
 $ SEXO     : Factor w/ 2 levels "hombre","mujer": 1 1 1 1 1 1 2 1 1 1 ...
```

```
$ EDAD     : Factor w/ 6 levels "menos de 19",..: 3 3 4 4 3 2 5 4 5 4 ...
$ ESTU     : Factor w/ 7 levels "1","2","3","4",..: 4 7 5 4 7 7 4 5 3 7 ...
$ ANTI     : Factor w/ 5 levels "Menos de 1 año",..: 5 4 3 5 4 2 5 5 5 2 ...
$ TIPOJOR  : Factor w/ 2 levels "Completa","Parcial": 1 1 1 1 1 1 1 1 1 1 ...
$ TIPOCON  : Factor w/ 2 levels "Indefinido","Temporal": 1 1 1 1 1 1 1 1 1 1 ...
$ RESPONSA : Factor w/ 2 levels "No","Sí": 2 2 2 2 2 2 1 1 1 1 ...
$ OCUPACION: Factor w/ 17 levels "A0","B0","C0",..: 1 3 4 4 4 4 5 4 4 4 ...
```

Como se muestra en la salida de la descripción de los datos, tenemos 209,436 observaciones y trabajaremos con 9 variables.

Para este capítulo se utilizan principalmente las referencias de las metodologías de EES y EPA, INE (2017).

Capítulo 3

Métodos estadísticos

3.1. Regresión logística

La regresión logística es una técnica analítica que nos permite relacionar funcionalmente una variable dicotómica con un conjunto de variables independientes, llamadas covariables, ya sean cualitativas o cuantitativas.

Esta técnica se enmarca en el conjunto de Modelos Lineales Generalizados (GLM, por sus siglas en inglés) en los que la distribución es binomial y la función de enlace es el logaritmo de las razones de probabilidad, es decir, la función *logit* que les da nombre.

Lo que buscamos es obtener la probabilidad de éxito a través del logaritmo de las razones de probabilidad de $Y = 1$, que obtenemos dividiendo las probabilidades 1 por la probabilidad de que sea 0. Así, estimamos por medio de Máxima Verosimilitud nuestra probabilidad desconocida, dada una determinada combinación lineal de variables independientes.

La variable dependiente será una variable dicotómica que se codificará como 0 o 1 ("ausencia" o "presencia", respectivamente).

El modelo logístico consiste en expresar la probabilidad de éxito de la siguiente manera:

$$P(Y = 1|X) = \frac{\exp(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)} \tag{3.1}$$

donde

$P(Y = 1|X)$ es la probabilidad de que Y tome el valor 1 (presencia de la característica estudiada).

X es un conjunto de n covariables $x_1, \ldots, x_n$ que forman parte del modelo.

β_0 es la constante del modelo o término independiente.

β_i los coeficientes de las covariables.

Algunas de las finalidades de los modelos de regresión logística son las siguientes:

- Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente.
- Clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente (es decir, los odds ratio para cada covariable).
- Clasificar individuos dentro de las categorías de la variable dependiente.

3.1.1. *Odds, odds ratio* (OR) y coeficientes

En una variable dicotómica Y se define la *odds* (disparidad o ventaja) como el cociente entre la probabilidad de éxito y la probabilidad de fracaso, esto es,

$$Odds(Y) = \frac{P(Y=1)}{P(Y=0)} = \frac{P(evento)}{1-P(evento)}$$

A partir de la probabilidad de éxito se puede averiguar el valor de la *odds*, pero también al revés. Son solo dos formas distintas de medir el parámetro desconocido de una variable dicotómica. Sin embargo, mientras la probabilidad de éxito toma valores en el intervalo [0, 1], la *odds* puede tomar cualquier valor real positivo, esto es, toma valores en el intervalo [0, ∞].

Como la *Odds* se mueve en el intervalo [0, ∞], solo falta aplicar un logaritmo para transformarlo en una cantidad situada en [$-\infty$, ∞], susceptible de ser explicada mediante un modelo lineal. En definitiva, el modelo consistirá en expresar el logaritmo de la *odds* de la variable respuesta como función lineal de la variable explicativa:

$$log\frac{\pi(x,\beta)}{1-\pi(x,\beta)}$$

Se invierte la función *logit* para adoptar el modelo como representación de la probabilidad de éxito. La inversa de la función responde a la expresión,

$$g^{-}1(x) = \frac{e^x}{1+e^x}$$

Finalmente, el modelo logístico consiste en expresar la probabilidad de éxito como expresamos previamente, en la ecuación 3.1.

Interpretación de los coeficientes de regresión

La *odds ratio* es una medida que juega un papel muy importante para la interpretación del modelo logístico, y en general en el estudio de variables dicotómicas.

Se define la *odds ratio* (cociente de disparidades) entre dos poblaciones respecto de una variable dicotómica Y como el cociente de las *odds* en una y otra población, esto es,

$$OddsRatio = \frac{Odds(Y/Pob2)}{Odds(Y/Pob1)} = \frac{P(Y=1/Pob2)/P(Y=0/Pob2)}{P(Y=1/Pob1)/P(Y=0/Pob1)}$$

Así, el parámetro β_1 resulta ser el logaritmo de la *odds ratio* entre las dos poblaciones: $X=0$ y $X=1$, y su estimador es el logaritmo de la *odds ratio* muestral. Por ejemplo,

$$Odss\widehat{R}atio = e^{\widehat{\beta}_1} = \frac{\widehat{p}_1/(1-\widehat{p}_1)}{\widehat{p}_0/(1-\widehat{p}_0)}$$

$$\beta_1 = log(Odss\widehat{R}atio)$$

Si el coeficiente es positivo se asocia con una *odds ratio* mayor que uno, que indica incremento de la *odds*. Si el coeficiente fuera negativo, la *odds ratio* sería menor que uno e indicaría una disminución de la *odds*.

Como regla general, los coeficientes de regresión pueden presentar signo negativo o positivo, mientras que la *odds* y la *odds ratio* son cantidades positivas, y se valora si son mayores o menores que uno.

La interpretación del *odds ratio* es que valores mayores que 1 indican que si el predictor aumenta, los *odds* de la variable dependiente crecen. Inversamente, un valor menor que 1 indica que tal como el predictor aumente el *odds* del resultado decrece.

3.2. Análisis de la varianza: interacción entre variables explicativas factoriales.

Los modelos factoriales de análisis de varianza sirven para evaluar el efecto individual y conjunto de dos o más factores sobre una variable dependiente cuantitativa. Utilizar más de un factor en un mismo diseño posee la ventaja de poder estudiar el efecto que la interacción entre ellos genera en la variable respuesta. En un modelo de dos factores, los efectos de interés son tres: los dos efectos principales (uno por cada factor) y el efecto de la interacción entre ambos factores.

En un análisis de varianza factorial existe una hipótesis nula por cada factor y por cada posible combinación de factores:

- La hipótesis nula referida a un factor afirma que las medias de las poblaciones definidas por los niveles del factor son iguales.
- La hipótesis referida al efecto de una interacción afirma que tal efecto es nulo.

Para contrastar estas hipótesis, el ANOVA factorial se sirve de estadísticos F así pues, para cada efecto existe una hipótesis y para cada hipótesis un estadístico F que permite contrastarla. El nivel crítico asociado a cada estadístico F es quien nos permite decidir si debemos mantener o rechazar una hipótesis.

El numerador de este estadístico es una estimación de la varianza poblacional basada en la variabilidad existente entre las medias de cada grupo: $\widehat{\sigma}_1^2 = n\widehat{\sigma}_{\bar{y}}^2$. El denominador del estadístico F es también una estimación de la varianza poblacional, pero basada en la variabilidad existente dentro de cada grupo: $\widehat{\sigma}_2^2 = \bar{S}_j^2$, donde j se refiere a los distintos grupos o niveles del factor:

$$F = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} = n\frac{\widehat{\sigma}_{\bar{y}}^2}{\bar{S}_j^2}$$

Si las medias poblacionales son iguales, las medias muestrales serán parecidas, existiendo entre ellas tan sólo diferencias atribuibles al azar. En ese caso, la estimación $\widehat{\sigma}_1^2$ basada en las diferencias entre las medias reflejará el mismo grado de variación que la basada en las diferencias entre las puntuaciones individuales $\widehat{\sigma}_2^2$ y el cociente F tomará un valor próximo a 1. Si las medias muestrales son distintas, la estimación $\widehat{\sigma}_1^2$ reflejará mayor grado de variación que la estimación $\widehat{\sigma}_2^2$ y el cociente F tomará un valor mayor que 1. Cuanto más diferentes sean las medias, mayor será el valor de F .

Si las poblaciones muestreadas son normales y sus varianzas son iguales, F se distribuye según el modelo de probabilidad F de *Fisher-Snedecor* con $(j-1)$ grados de libertad del numerador y $(n-j)$ del denominador, donde n es el número total de observaciones. Si suponemos cierta la hipótesis de igualdad de medias, podemos conocer en todo momento la probabilidad de obtener un valor como el obtenido o mayor

El estadístico F se interpreta de forma que si su nivel crítico asociado, es decir, si la probabilidad de obtener valores como el obtenido o mayores, es menor que 0,05 rechazaremos la hipótesis de igualdad de medias y concluiremos que no todas las medias poblacionales comparadas son iguales. En caso contrario, no podremos rechazar la hipótesis de igualdad y no podremos afirmar que los grupos comparados difieran en sus promedios poblacionales.

La interacción entre variables explicativas factoriales se refiere a la interacción entre dos variables factoriales, por ejemplo A y B. Del factor A podemos encontrar I valores posibles, mientras que el factor B admite J posibilidades. Nos interesa determinar el efecto de cada uno de estos dos factores, y también su interacción.

Por tanto, vamos a pensar que además del efecto que cada uno puede ejercer por separado, al juntarse ciertos valores del factor A con ciertos valores del factor B, el efecto total que producen no es sólo la suma de los efectos de cada uno, sino que se puede producir una interacción.

En cada una de las IxJ posibilidades realizaremos varias observaciones de la variable respuesta Y. Si se realiza el mismo número K observaciones en cada combinación de los dos factores, entonces diremos que el diseño es equilibrado, lo cual presenta ciertas ventajas en los procedimientos que se vayan a aplicar. En cualquier caso, en términos generales podemos suponer que se toman n_{ij} observaciones para el nivel i del factor A y el nivel j del factor B.

Si se adopta una parametrización por desviación respecto de la media global, el modelo quedará planteado de la siguiente manera:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \qquad i \in \{1, \ldots, I\} \quad j \in \{1, \ldots, J\} \quad k \in \{1, \ldots, n_{ij}\}$$

donde

$\varepsilon_{ijk} \in N(0, \sigma^2)$ y todos ellos independientes entre sí.

El parámetro μ representa la media global.

Los parámetros α_i representa el efecto del nivel i del factor A.

Los parámetros β_j representa el efecto del nivel j del factor B.

Los parámetros γ_{ij} representan la interacción de los factores A y B.

ε_{ijk} error experimental.

Si se adopta la parametrización por desviación respecto de unas categorías de referencia, el modelo se expresará así:

$$Y_{ijk} = \mu_{11} + \alpha_i I_{i\neq 1} + \beta_j I_{i\neq 1} + \gamma_{ij} I_{i\neq 1} I_{j\neq 1} + \varepsilon_{ijk} \qquad i \in \{1, \ldots, I\} \quad j \in \{1, \ldots, J\} \quad k \in \{1, \ldots, n_{ij}\}$$

siendo

$\varepsilon_{ijk} \in N(0, \sigma^2)$ y todos ellos independientes entre sí.

El parámetro μ_{11} representa la media de Y para la categoría 1 del factor A y la categoría 1 del factor B.

Los parámetros α_i representan los efectos principales de desviación respecto de la primera categoría del factor A (la notación $I_{i\neq 1}$ indica que sólo se toman estos parámetros si la categoría no es la de referencia). Con los parámetros β_j ocurre algo similar como desviaciones respecto de la primera categoría del factor B.

Los parámetros γ_{ij} representan la interacción de los factores A y B, pero suprimiendo la primera categoría tanto del factor A como del B. Como resultado se tienen $(I-1)$ x $(J-1)$ parámetros de interacción.

Este capítulo se ha realizado con la ayuda de Montgomery (2013), el capítulo 8 de Sheather (2009), y Hosmer y Lemeshow (1989).

Capítulo 4

Resultados: Datos EPA

Mediante el estudio estadístico descriptivo y la aplicación de un modelo de regresión logística se pretende describir y conocer qué condiciones favorecen o no la aparición del desempleo, identificando los factores de riesgo, entre las variables disponibles, para la situación de parado.

4.1. Análisis exploratorio

En esta parte se presentan una serie de gráficos y medidas propias para cada variable estudiada. Se aplican estadísticos de asociación de variables, todo esto con el objetivo de conocer la relación entre los factores.

En la Tabla 4.1 se muestra la cantidad de ocupados y parados, según las definiciones antes explicadas extraídas desde la página del INE.

Cuadro 4.1: Número de activos.

ACTIVOS	n
ocupados	64635
parados	10227

4.1.1. Activos según datos demográficos y formación

Activos en función del sexo

Al analizar los `Activos` de los datos EPA 2019-T3 según la variable `Sexo`, se observa que el mayor porcentaje de los parados son mujeres con un 54.83 %. Mientras que en el grupo de ocupados estas representan el 46.83 %.

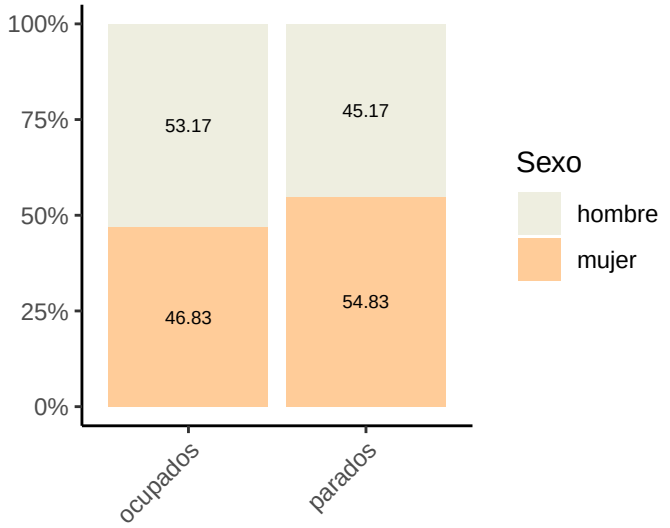


Figura 4.1. Porcentaje de activos por sexo.

Se aplica el *test* χ^2 de independencia, también conocido como χ^2 de *Pearson* que se emplea para estudiar si existe asociación entre dos variables categóricas, es decir, si las proporciones de una variable son diferentes dependiendo del valor que adquiera la otra variable, cuando los datos son independientes.

La formulación de las hipótesis son las siguientes:

H_0: Las variables son independientes por lo que una variable no varía entre los distintos niveles de la otra variable.

H_a: Las variables son dependientes, una variable varía entre los distintos niveles de la otra variable.

Su fórmula estadística esta dada por:

$$\chi^2 = \sum_{i,j} \frac{(observado_{i,j} - esperado_{i,j})^2}{esperado_{i,j}}$$

El valor esperado de cada grupo se obtiene multiplicando las frecuencias marginales de la fila y columna en la que se encuentra la celda y dividiendo por el total de observaciones. Se suman las diferencias de todos los niveles. Elevar al cuadrado las diferencias permite hacerlas todas positivas y permite además magnificar aquellas más grandes.

Con la función `chisq.test()` aplicamos el estadístico en R. Se obtiene un *p-value* por debajo del nivel de significación de 0.05, rechazando la hipótesis nula de independencia entre las variables.

```
tabla_act_sex <- table(ACTIVOS, SEXO)
chisq.test(tabla_act_sex)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

  data:  tabla_act_sex
  X-squared = 226, df = 1, p-value <2e-16
```

Dado que el *test* contrasta si las variables están relacionadas, al tamaño del efecto se le conoce como fuerza de asociación. Existen múltiples medidas de asociación, entre las que destacan: el `coeficiente phi` es el valor de la `chi cuadrado` entre el número de observaciones, un valor próximo a 0 indica independencia entre los factores, valores próximos o superiores a 1 implican relación entre los factores. El `coeficiente de contingencia` también es una medida de la intensidad de la relación basado en la `chi cuadrado` y toma valores entre 0 (independencia) y 1 (dependencia). La `V de Cramer` es muy habitual para medir la relación entre factores, es menos susceptible a valores muestrales. También 0 implica independencia y 1 una relación perfecta entre los factores.

En R se pueden calcular estas medidas mediante la función `assocstats()` del paquete `vcd`.

```
assocstats(tabla_act_sex)
```

```
                    X^2 df P(> X^2)
Likelihood Ratio 225.95  1        0
Pearson          226.05  1        0

Phi-Coefficient   : 0.055
Contingency Coeff.: 0.055
Cramer's V        : 0.055
```

En este caso tenemos valores muy próximos a 0 en todos los estadísticos que nos ofrece, concluyendo que ambos factores tienen un nivel de asociación bajo.

Activos por grupos de edad

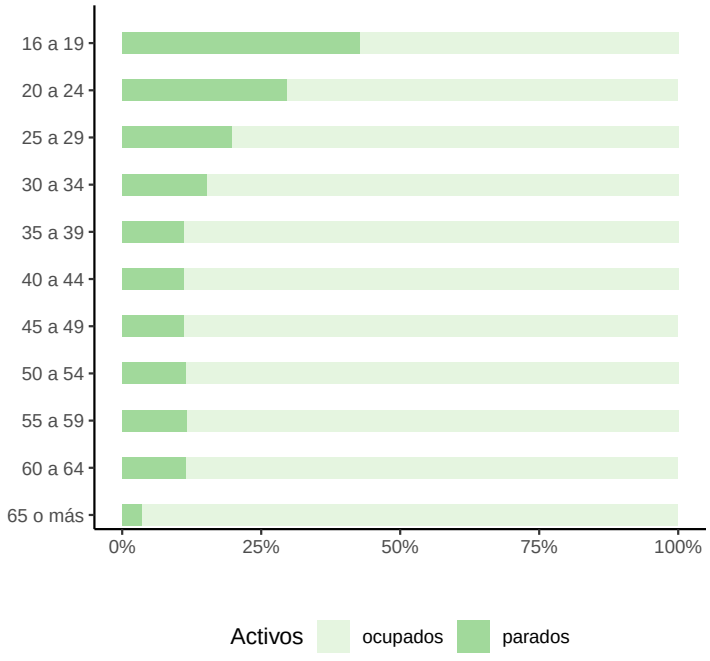


Figura 4.2. Porcentaje de activos por edad.

Cuadro 4.2: Activos por grupos de edad.

EDAD	ocupados_n	ocupados_porc	parados_n	parados_porc
16 a 19	612	57.25	457	42.75
20 a 24	2971	70.40	1249	29.60
25 a 29	4495	80.24	1107	19.76
30 a 34	5575	84.77	1002	15.23
35 a 39	7676	88.85	963	11.15
40 a 44	9890	88.89	1236	11.11
45 a 49	9919	88.98	1229	11.02
50 a 54	9721	88.50	1263	11.50
55 a 59	8221	88.32	1087	11.68
60 a 64	4718	88.65	604	11.35
65 o más	837	96.54	30	3.46

En la Figura 4.2 se observa que entre los mas jóvenes el porcentaje de parados es mayor.

Se rechaza la hipótesis independencia de las variables al aplicar el estadístico `chi cuadrado`. Concluimos que existe una relación entre la `Edad` y la situación en que se encuentra la persona en el mercado laboral. Aun así el nivel de asociación es bajo.

```
  Pearson's Chi-squared test

data:  tabla_act_edad
X-squared = 2215, df = 10, p-value <2e-16
```

```
                    X^2 df P(> X^2)
Likelihood Ratio 1822.2 10        0
Pearson          2214.7 10        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.17
Cramer's V        : 0.172
```

Activos en función del estado civil

En el caso de los ocupados predominan los casados con 57.11 %, mientras que en el grupo de parados predominan con el 50.50 % los solteros.

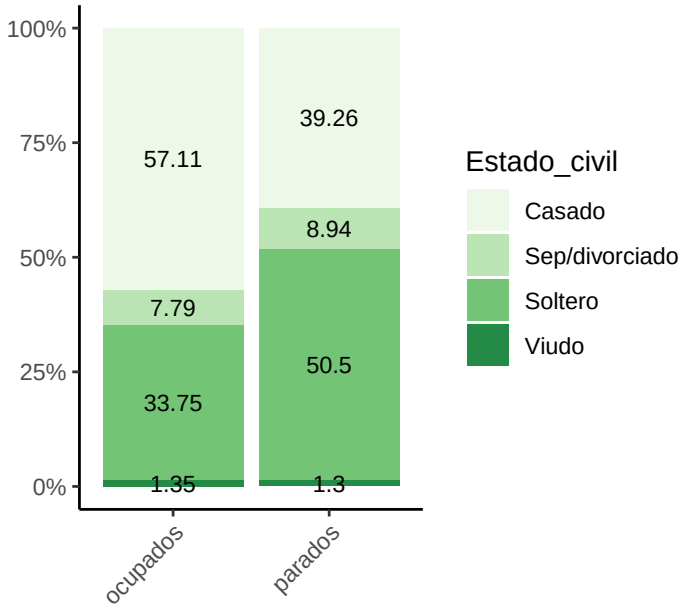


Figura 4.3. Activos por estado civil.

```
  Pearson's Chi-squared test

data:  t_activos_ecivi
X-squared = 1218, df = 3, p-value <2e-16
```

Con los estadísticos se comprueba que hay una asociación baja entre el `Estado civil` y su situación en el mercado laboral.

```
                    X^2 df P(> X^2)
Likelihood Ratio 1200.4  3        0
Pearson          1217.5  3        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.127
Cramer's V        : 0.128
```

Activos por comunidad autónoma

Son Ceuta, Melilla, seguido por Andalucía y Canarias las `Comunidades autónomas` con un mayor porcentaje de parados. Mientras el mayor porcentaje de ocupados se encuentran en Navarra, Illes Baleares y País Vasco.

Al aplicar el estadístico `chi cuadrado` observamos que se rechaza la hipótesis nula de independencia de las variables. El resultado de los estadístico de asociación indican que los factores tienen un nivel bajo de relación.

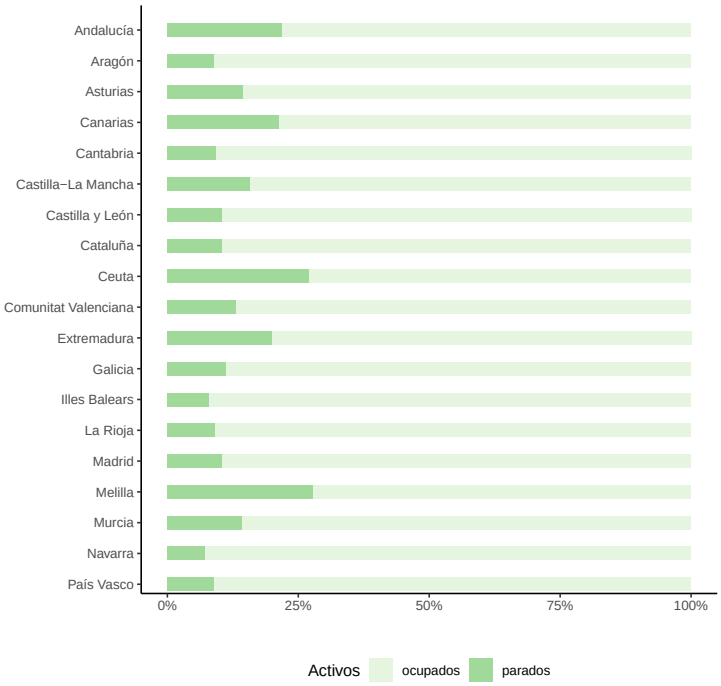


Figura 4.4. Activos por comunidad autónoma.

Cuadro 4.3: Activos por comunidad autónoma.

CCAA	ocupados_n	ocupados_por	parados_n	parados_por
Andalucía	9261	78.22	2578	21.78
Aragón	3043	91.03	300	8.97
Asturias	1693	85.68	283	14.32
Canarias	2434	78.80	655	21.20
Cantabria	1533	90.71	157	9.29
Castilla-La Mancha	4336	84.29	808	15.71
Castilla y León	6166	89.57	718	10.43
Cataluña	7139	89.64	825	10.36
Ceuta	230	73.02	85	26.98
Comunitat Valenciana	4997	86.84	757	13.16
Extremadura	1939	80.02	484	19.98
Galicia	7591	88.78	959	11.22
Illes Balears	1951	91.98	170	8.02
La Rioja	1186	91.02	117	8.98
Madrid	4073	89.60	473	10.40
Melilla	216	72.24	83	27.76
Murcia	2108	85.87	347	14.13
Navarra	1571	92.90	120	7.10
País Vasco	3168	91.14	308	8.86

```
    Pearson's Chi-squared test

data:  tabla_act_ccaa
X-squared = 1530, df = 18, p-value <2e-16

                    X^2 df P(> X^2)
Likelihood Ratio 1456.0 18        0
Pearson          1529.9 18        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.142
Cramer's V        : 0.143
```

Activos por tipo de nacionalidad

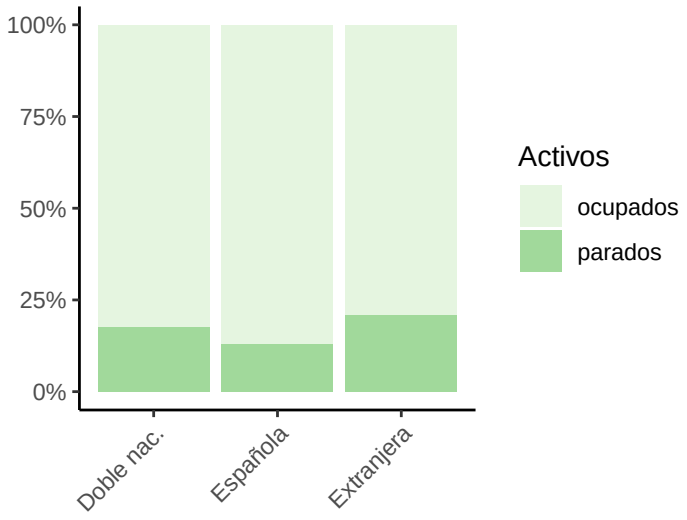


Figura 4.5. Activos por tipo de nacionalidad.

Cuadro 4.4: Activos por nacionalidad.

NAC	ocupados_n	ocupados_por	parados_n	parados_por
Doble nac.	1863	82.47	396	17.53
Española	58379	87.07	8667	12.93
Extranjera	4393	79.05	1164	20.95

Al relacionar la variable `Nacionalidad` con la situación de empleo se constata que existe una asociación baja entre las variables.

```
    Pearson's Chi-squared test

data:  tabla_act_nac
X-squared = 309, df = 2, p-value <2e-16
```

```
                   X^2 df P(> X^2)
Likelihood Ratio 278.28  2        0
Pearson          309.37  2        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.064
Cramer's V        : 0.064
```

Activos por nivel de estudio

En esta variable clasifican a la población mayor o igual a 16 años de la siguiente manera:

Cuadro 4.5: Clasificación por nivel de estudio.

Cód.	Descripción
AN	Analfabetos
P1	Educación primaria incompleta
P2	Educación primaria
S1	Primera etapa de educación secundaria
SG	Segunda etapa de educación secundaria. Orientación general
SP	Segunda etapa de educación secundaria. Orientación profesional
SU	Educación superior

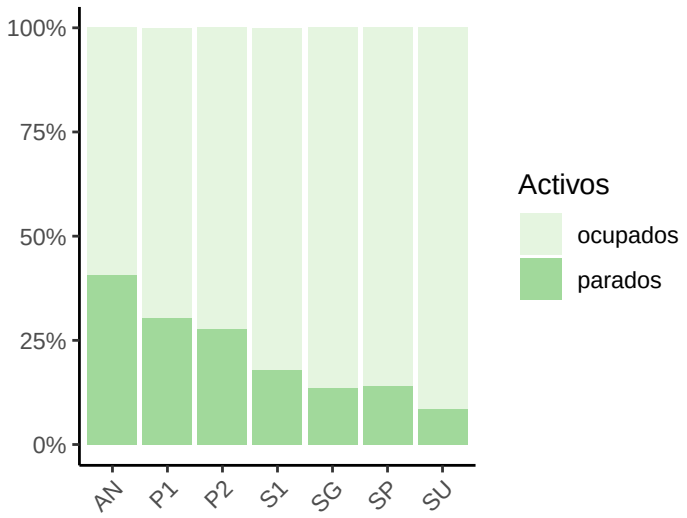


Figura 4.6. Activos por tipo de estudio.

Hay un gran número de los entrevistados concentrados en el grupo de "Primera etapa de educación secundaria" y "Educación superior". En el primer caso los ocupados conforman un 82.21 %, mientras que en el segundo conforman el 91.48 %.

Si verificamos la Tabla 4.6 y la Figura 4.6, a medida que aumenta la formación se incrementa la probabilidad de pertenecer a la población ocupada.

Al relacionar la variable de `Nivel de estudios` con la situación de empleo se verifica que existe una asociación baja entre las variables.

Cuadro 4.6: Activos por nivel de estudio.

ESTU	ocupados_n	ocupados_por	parados_n	parados_por
AN	95	59.38	65	40.62
P1	475	69.55	208	30.45
P2	2563	72.30	982	27.70
S1	18088	82.21	3914	17.79
SG	8436	86.45	1322	13.55
SP	6791	85.95	1110	14.05
SU	28187	91.48	2626	8.52

```
  Pearson's Chi-squared test

data:  tabla_act_estu[, c(-1)]
X-squared = 1863, df = 6, p-value <2e-16

                    X^2 df P(> X^2)
Likelihood Ratio 1753.9  6        0
Pearson          1863.2  6        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.156
Cramer's V        : 0.158
```

Análisis de correspondencia múltiple

Para estudiar la relación entre la actividad económica y las variables analizadas previamente, de forma conjunta, se aplica un análisis de correspondencia múltiple para una mejor ilustración e interpretación.

El análisis de correspondencias es una técnica estadística que se utiliza para analizar, desde un punto de vista gráfico, las relaciones de dependencia e independencia de un conjunto de variables categóricas a partir de los datos de una tabla de contingencia.

Para ello asocia a cada una de las modalidades de la tabla, un punto en el espacio R^n, (generalmente=2), de forma que las relaciones de cercanía o lejanía entre los puntos calculados reflejen las relaciones de dependencia y semejanza existentes entre ellas.

Este representa los perfiles de fila y columna simultáneamente en un espacio común Bendixen (2003). Los individuos más próximos tienen patrones similares. Asimismo, los individuos más cercanos a un determinado perfil de columna, indica que en su perfil tiene un porcentaje alto de esa categoría.

La extensión del análisis de correspondencias simples al caso de más de dos variables nominales (tablas de contingencia multidimensionales) se denomina "Análisis de Correspondencias Múltiples", y utiliza los mismos principios generales que la técnica simple. En general, se orienta a casos en los cuales una variable representa *ítems* o individuos y el resto son variables cualitativas u ordinales que representan cualidades.

Para implementar el análisis utilizamos la función `MCA()` de la librería `FactoMineR`.

```
res.mca <- MCA(mac_activos, graph = FALSE)
```

En la Figura 4.7, se grafican los resultado para conocer la asociación de las categorías de las variables. Se realiza el gráfico representando dos primeras dimensiones. Las categorías tendrían un color en función de su

calidad, seleccionando el parámetro `col.var`="cos2" en la función `fviz_mca_var()`. Los cosenos al cuadrado permiten saber si un punto está bien representado sobre el eje factorial. La calidad de la representación de un punto sobre el eje será tanto mayor cuando más próximo a 1 sea el coseno al cuadrado.

Seleccionamos las 35 categorías que más han contribuido a las dimensiones 1 y 2 para una mejor ilustración.

```
fviz_mca_var(res.mca, col.var = "cos2", select.var = list(contrib = 35),
    repel = TRUE, gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
    ggtheme = theme_minimal())
```

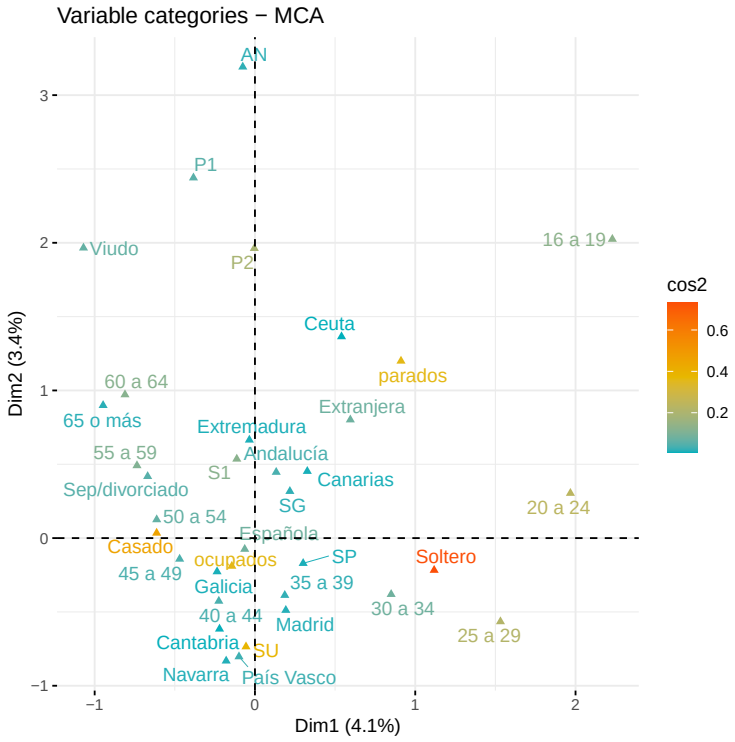


Figura 4.7. Relación de las categorías de variables-MCA.

Tomando en cuenta que la proximidad entre categorías de variables se interpreta en términos de asociación o dependencia, las características principales asociadas a las actividades económicas son las siguientes:

- Los "ocupados" están caracterizados principalmente por ser personas de `Edades` entre "35 y 49" años, con `Niveles de estudios` de "Educación superior" y "Segunda etapa de educación secundaria. Orientación profesional", y con `Estado civil` "casado" y `Nacionalidad` "española". Relacionados a `Comunidades autónomas` como "Galicia", "Cantabria", "Navarra" y "País Vasco" y "Madrid".
- Los "parados" estarían caracterizados principalmente por ser "jóvenes", con "Educación secundaria", "extranjeros". Las comunidades "Ceuta", "Canarias", "Andalucía", "Extremadura" son las más cercanas a este grupo.

La variable `Sexo`, algunas `Comunidades autónomas`, la categoría "Doble nac" entre otras, resultan un poco más difícil clasificarlas en un grupo de actividad económica. En algunos casos tienen una calidad de representación baja, otras ya no salen representadas, porque limitamos la cantidad de categorías en 35. Es importante señalar que las variables de color más anaranjado son las que más han contribuido a explicar las relaciones entre las variables, mientras que las de color azul más intenso las que menos han contribuido.

4.1.2. Ocupados según situación profesional

Interesa conocer cuál es la situación de los ocupados.

Ocupación principal

Cuadro 4.7: Ocupación principal.

OCUP	n	Porcentaje
Trabajadores de serv. de restauración	14532	22.48
Técn. y Profesionales científicos	11771	18.21
Trabajadores no cualificados	7746	11.98
Trabajadores cualificados manuf. y const.	7298	11.29
Empleados administrativo	6686	10.34
Técn. y Profesionales de apoyo	6641	10.27
Operadores de instalación/maquinaria	5147	7.96
Directores y gerentes	2580	3.99
Trabajadores cualificados agric. y pesca	1859	2.88
Ocupaciones militares	375	0.58

El 22.48 % de los ocupados se encuentran en el grupo de "Trabajadores de servicios de restauración, personales, protección y vendedores de comercio", seguido por los "Técnicos y Profesionales científicos e intelectuales", los de menor frecuencia son los de "ocupaciones militares" con tan solo un 0.58 %.

Cuál es su situación profesional (actividad principal)?

Cuadro 4.8: Actividad principal de los ocupados.

SITU_PROF	n	Porcentaje
Asalariado sector privado	41417	64.08
Asalariado sector público	12343	19.10
Empresario con asalariados	3492	5.40
Miembro de una cooperativa	103	0.16
Negocio familiar	265	0.41
Otra situación	18	0.03
Trabajador independiente	6997	10.83

De todos los encuestados ocupados el 83.18 % son asalariados, de estos el 64.08 % son "asalariados del sector privado" y el 19.10 % "asalariados del sector público".

Tipo de contrato de los asalariados

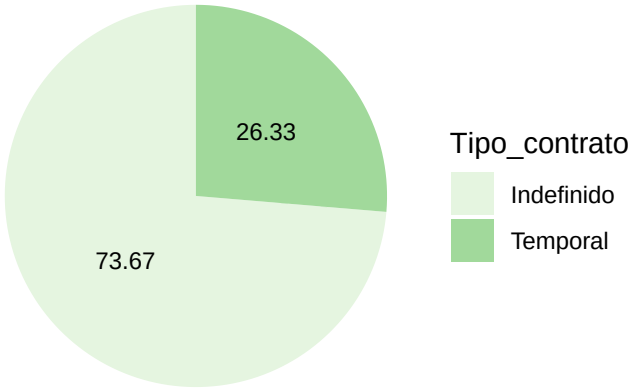


Figura 4.8. Ocupados por tipo de contrato.

De los ocupados asalariados el 73.67 % tienen un contrato indefinido.

Al analizar el `Tipo de contrato` en función del `Sexo`, son los hombres quienes tienen un mayor porcentaje de contratos indefinidos, mientras que las mujeres tienen un mayor porcentaje de contratos temporales. Sin embargo, la diferencia en porcentajes no parece ser significativa.

Cuadro 4.9: Tipo de contrato en función del sexo.

TIPOCON	hombre_n	hombre_porc	mujer_n	mujer_porc
Indefinido	20485	51.72	19121	48.28
Temporal	6848	48.38	7306	51.62

Tipo de jornada (de todos los ocupados)

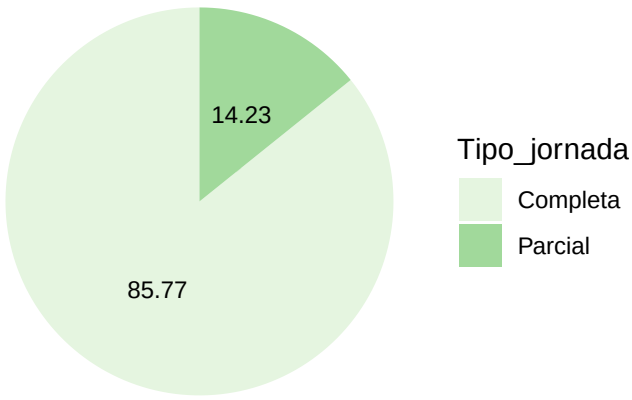


Figura 4.9. Ocupados por tipo jornada.

El 85 % de los encuestados clasificados como ocupados tienen una jornada completa.

Es importante destacar las diferencia existente entre los hombres y mujeres con respecto a la jornada de trabajo, especialmente en la jornada parcial donde el 76.2 % son mujeres y solo un 42 % de ellas tiene un contrato de jornada completa.

Cuadro 4.10: Tipo de jornada en función del sexo.

TIPOJOR	hombre_n	hombre_porc	mujer_n	mujer_porc
Completa	32173	58.03	23265	41.97
Parcial	2192	23.83	7005	76.17

Motivo de tener jornada parcial

Al preguntarles el motivo por el cuál tienen una jornada parcial, de los 9,197 el 50.7 % responde no haber podido encontrar un trabajo de jornada completa, seguido por el 12.69 % que especifica que están al cuidado de niños o adultos enfermos, solo el 10.7 % dice no querer un trabajo de jornada completa.

Cuadro 4.11: Motivo jornada parcial.

MOT_JP	n	porcentaje
No encuentra jornada completa	4663	50.70
Cuidado de niños/adultos enfermos	1167	12.69
Otras razones	997	10.84
No quiere jornada completa	975	10.60
Otras respons. familiares/personales	647	7.03
Cursando estudio/formación	569	6.19
Enfermedad/ incapacidad propia	132	1.44
Desconoce el motivo	47	0.51

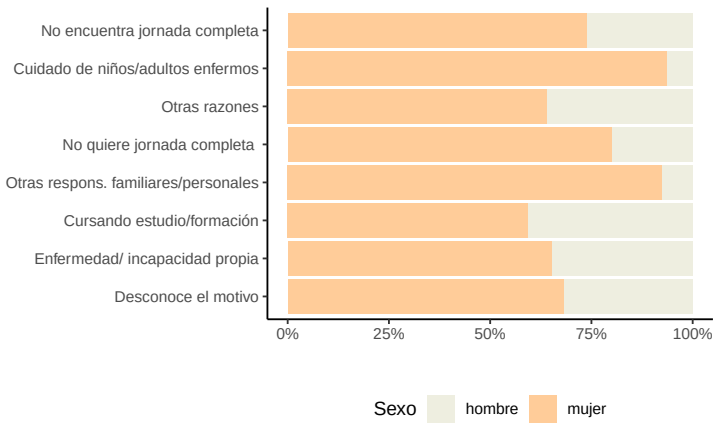


Figura 4.10. Motivo jornada parcial por sexo.

Es importante destacar que las mujeres son las que tienen mayor responsabilidad cuando de los temas familiares se trata. En los casos donde responden que el `Motivo de tener jornada parcial` es el "Cuidado de niños/adultos enfermos" y "Otras responsabilidades familiares/personales" más del 90 % son mujeres.

4.2. Inferencia

Para continuar con el estudio interesa conocer cuáles son los factores de riesgo, entre las variables disponibles, para la situación de parado. Aquellas circunstancias o situaciones que aumentan las probabilidades de estar parado.

Interesa estudiar la cuantificación de estos factores de riesgo y para ello procedemos a aplicar regresión logística que permite considerar varios factores simultáneamente y calcular *odds ratios* que miden el incremento de riesgo.

Variables del modelo

- La variable a explicar: `Activos`.

Con los valores 0 para los ocupados, que será la categoría de referencia y 1 para los que están parados, que será nuestra categoría de estudio.

- Variables explicativas:

Se escogen las variables indicadas en la Tabla 4.12, relacionadas con la actividad económica que de acuerdo al análisis descriptivo antes realizado pensamos que podrían tener un efecto sobre la probabilidad de estar en la condición de parado frente a encontrarse ocupado.

Ajuste del modelo

La función básica para regresión logística, y en general, para modelo lineales generalizados, que son los modelos que se convierten en lineales a través de una función *link*, es la función `glm()`.

La forma de la función es:

$$glm(dependiente \sim independiente1 + independiente2 +, family = binomial(), data = datos)$$

El primer argumento es un objeto de la clase fórmula. A la izquierda del signo ~ ubicamos a la variable dependiente y a la derecha, unidos por el signo + las independientes. El segundo, family=binomial(), especifica la función de probabilidad que utilizaremos. Para modelos *logit* es una función binomial. Dentro de los paréntesis se puede especificar la función de enlace. Para la familia de distribuciones binomial `glm()` por defecto usa una función *logit*.

En el modelo de regresión logística solo podemos introducir variables dicotómicas, entonces como las variables explicativas son categóricas y algunas tienen más de dos categorías R crea variables ficticias a partir de cada variable para poder usarla en el modelo.

Respecto de la parametrización, en el lenguaje R, el algoritmo de la función asume por defecto una categoría de referencia (la primera de la variable, generalmente por orden alfabético) como forma de parametrizar una variable explicativa factorial.

Los grupos de referencia son los siguientes:

- Hombre
- 16-19 años
- Soltero
- Analfabeto
- De nacionalidad española
- De Madrid [1]

[1] Único caso en el que cambiamos la categoría de referencia para facilitar la interpretación.

Cuadro 4.12: Variables explicativas

Variable	Cód	Descripción
Sexo	0	Hombre
	1	Mujer
Grupos de edad	16	16 a 19
	20	20 a 24
	25	25 a 29
	30	30 a 34
	35	35 a 39
	40	40 a 44
	45	45 a 49
	50	50 a 54
	55	55 a 59
	60	60 a 64
	65	65 o más
Estado civil	1	Soltero
	2	Casado
	3	Viudo
	4	Separado/divorciado
Nacionalidad	1	Española
	2	Española y doble nacionalidad
	3	Extranjera
Comunidad autónoma	1	Andalucía
	2	Aragón
	3	Asturias, Principado de
	4	Balears, Illes
	5	Canarias
	6	Cantabria
	7	Castilla y León
	8	Castilla-La Mancha
	9	Cataluña
	10	Comunitat Valenciana
	11	Extremadura
	12	Galicia
	13	Madrid, Comunidad de
	14	Murcia, Región de
	15	Navarra, Comunidad Foral de
	16	País Vasco
	17	Rioja, La
	51	Ceuta
	52	Melilla

Note:
Para el Nivel de estudio : Ver Tabla 4.5.

Todos los casos se enfrentan con la variable de referencia. Por ejemplo, en el caso de la variable `Edad`, como vemos en la salida del resumen del modelo, el algoritmo de la función ha creado las variables, todas ellas variables ficticias que muestran la diferencia entre cada grupo y el grupo de referencia, que en este caso es 16-19. Así, el valor para, por ejemplo, el grupo de 20-24, indica la diferencia del *logit* de encontrarse sin trabajo una persona del grupo de 16-19 años a una que se encuentre en el grupo de 20-24.

A continuación, se muestran los resultados del ajuste de un modelo de regresión logística sobre las variables explicativas. Nos fijamos en los coeficientes, sus errores estándar, el estadístico z y los p-valores asociados. En lugar de los resultados de coeficiente de determinación y *test* F, propios del modelo lineal, se presentan aquí los valores de la *deviance* y el número de iteraciones requeridas para obtener las estimaciones.

Este coeficiente en la regresión logística se puede interpretar como el cambio en el *logit* de la variable de resultado asociado al cambio de una clase a otra en la variable predictora, donde, como habíamos mencionado, el *logit* es simplemente el logaritmo natural de las probabilidades de que Y ocurra.

Un estadístico crucial es el estadístico de *Wald* (*z-statistic*) que tiene una distribución normal y nos dice si el coeficiente para ese predictor es significativamente diferente de cero para poder así suponer que la variable predictora está haciendo una contribución significativa a la predicción del resultado (Y).

```
modelo_glm <- glm(ACTIVOS ~ SEXO + EDAD + ECIVIL + ESTU + NAC + CCAA_,
    family = binomial(link = "logit"))
```

```
summary(modelo_glm)
```

```
Call:
glm(formula = ACTIVOS ~ SEXO + EDAD + ECIVIL + ESTU + NAC + CCAA_,
    family = binomial(link = "logit"))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.776  -0.562  -0.421  -0.310   2.977

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5299     0.1899    2.79   0.0053 **
SEXO1         0.4803     0.0229   20.98  < 2e-16 ***
EDAD20       -0.3748     0.0736   -5.09  3.5e-07 ***
EDAD25       -0.8395     0.0740  -11.34  < 2e-16 ***
EDAD30       -1.1081     0.0746  -14.85  < 2e-16 ***
EDAD35       -1.3741     0.0753  -18.24  < 2e-16 ***
EDAD40       -1.2643     0.0745  -16.97  < 2e-16 ***
EDAD45       -1.2883     0.0753  -17.11  < 2e-16 ***
EDAD50       -1.2428     0.0755  -16.45  < 2e-16 ***
EDAD55       -1.2396     0.0770  -16.09  < 2e-16 ***
EDAD60       -1.3499     0.0830  -16.26  < 2e-16 ***
EDAD65       -2.6632     0.2020  -13.18  < 2e-16 ***
ECIVIL2      -0.5672     0.0296  -19.14  < 2e-16 ***
ECIVIL3      -0.4507     0.1014   -4.44  8.8e-06 ***
ECIVIL4      -0.1074     0.0454   -2.37   0.0179 *
ESTUP1       -0.4615     0.1891   -2.44   0.0147 *
ESTUP2       -0.6751     0.1732   -3.90  9.7e-05 ***
ESTUS1       -1.1927     0.1703   -7.00  2.5e-12 ***
ESTUSG       -1.6569     0.1719   -9.64  < 2e-16 ***
ESTUSP       -1.5960     0.1730   -9.23  < 2e-16 ***
```

```
ESTUSU       -2.0550     0.1710  -12.02  < 2e-16 ***
NAC2          0.3636     0.0603    6.03  1.6e-09 ***
NAC3          0.5331     0.0390   13.66  < 2e-16 ***
CCAA_1        0.8600     0.0561   15.32  < 2e-16 ***
CCAA_2       -0.1746     0.0800   -2.18   0.0292 *
CCAA_3        0.5202     0.0833    6.24  4.3e-10 ***
CCAA_4       -0.5271     0.0965   -5.46  4.7e-08 ***
CCAA_5        0.7832     0.0682   11.48  < 2e-16 ***
CCAA_6       -0.0238     0.0998   -0.24   0.8112
CCAA_7        0.0426     0.0650    0.66   0.5118
CCAA_8        0.4296     0.0645    6.66  2.8e-11 ***
CCAA_9       -0.1248     0.0632   -1.98   0.0482 *
CCAA_10       0.2618     0.0647    4.04  5.3e-05 ***
CCAA_11       0.7236     0.0737    9.82  < 2e-16 ***
CCAA_12       0.1369     0.0617    2.22   0.0265 *
CCAA_14       0.1772     0.0790    2.24   0.0249 *
CCAA_15      -0.3352     0.1093   -3.07   0.0022 **
CCAA_16      -0.0887     0.0795   -1.12   0.2648
CCAA_17      -0.2226     0.1119   -1.99   0.0466 *
CCAA_51       1.0103     0.1430    7.07  1.6e-12 ***
CCAA_52       1.1494     0.1476    7.79  6.8e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 59705  on 74861  degrees of freedom
Residual deviance: 53884  on 74821  degrees of freedom
AIC: 53966

Number of Fisher Scoring iterations: 5
```

Debajo se muestran también las exponenciales de los coeficientes, que constituyen las OR para el efecto de cada una de las variables, fijados los valores de las demás. Se efectúa la exponencial de los coeficientes con el propósito de facilitar la interpretación de los coeficientes.

```
exp(coef(modelo_glm))
```

```
(Intercept)       SEXO1      EDAD20      EDAD25      EDAD30      EDAD35
    1.69877     1.61651     0.68741     0.43195     0.33018     0.25307
     EDAD40      EDAD45      EDAD50      EDAD55      EDAD60      EDAD65
    0.28243     0.27573     0.28859     0.28949     0.25926     0.06973
    ECIVIL2     ECIVIL3     ECIVIL4      ESTUP1      ESTUP2      ESTUS1
    0.56713     0.63721     0.89814     0.63031     0.50908     0.30339
     ESTUSG      ESTUSP      ESTUSU        NAC2        NAC3      CCAA_1
    0.19074     0.20271     0.12810     1.43845     1.70413     2.36314
     CCAA_2      CCAA_3      CCAA_4      CCAA_5      CCAA_6      CCAA_7
    0.83980     1.68237     0.59033     2.18853     0.97644     1.04354
     CCAA_8      CCAA_9     CCAA_10     CCAA_11     CCAA_12     CCAA_14
    1.53664     0.88271     1.29922     2.06181     1.14675     1.19383
    CCAA_15     CCAA_16     CCAA_17     CCAA_51     CCAA_52
    0.71519     0.91513     0.80041     2.74636     3.15615
```

Al observar el coeficiente de la variable `Sexo` interpretamos que las mujeres tienen una mayor probabilidad de encontrarse en la condición de parado. Observamos una *odds ratio* de 1.61 por lo que la *odds* de encontrarse parados se incrementa en un 61 %.

De 16-19 años es el grupo de referencia de la variable `Edad`. Pasar de este grupo a los demás disminuye la probabilidad de encontrarse en la situación de parado o mejor dicho encontrarse en grupos de edades mayores disminuye la probabilidad de estar parado. En unos casos más que otro y lo confirmamos con la exponencial.

Para facilitar su interpretación calculamos la inversa para conocer el riesgo (1/exp). Entonces, en este caso a mayor valor mayor riesgo tienen los del grupo de referencia frente a los demás grupos a estar parados. Si aplicamos la inversa el mayor riesgo lo presentan frente a edades del grupo de 65 o más y el menor riesgo con las edades más cercanas a ellos que son los de 20-24.

Con respecto al `Estado civil`, en todos los casos tenemos valores de *odds ratio* por debajo de 1, por lo que estar soltero representa un factor de riesgo frente a estar casado, separado o viudo.

Como era de esperar, pasar del `Nivel de formación` más bajo (analfabeto, que es el grupo de referencia) a los niveles más altos disminuye la probabilidad de estar sin trabajo. Si aplicamos la inversa a los exponentes cuando comparamos el riesgo de una persona analfabeta de estar parado frente a los demás niveles de formación, este riesgo aumenta mientras aumenta el nivel de formación.

Con respecto a la `Nacionalidad`, en ambas categorías los coeficientes son positivos, por lo que el tener nacionalidad extranjera o tener doble nacionalidad aumenta la probabilidad de estar parado. Observamos una OR de 1.43 para los de doble nacionalidad y 1.70 para los extranjeros, por lo que la *odds* de estar parado se incrementa en un 43 % y a 70 % respectivamente para estas categoría.

En cuanto a la `Comunidad autónoma` tomamos como referencia a Madrid, en los únicos caso que la probabilidad de estar parado no aumenta y que son categorías significativas es cuando se trata de ciudades como: Aragón, Balears, Illes y Navarra. Algunos casos como: Cataluña, País Vasco y Cantabria, disminuye la probabilidad pero sus p-valores están por debajo del nivel de significación del 5 %, que entendemos que entre su clasificación de parados y ocupados no hay muchas diferencias.

En los demás aumenta la probabilidad, especialmente para Melilla y Ceuta.

Bondad de ajuste y evaluación del modelo

A la hora de evaluar la validez y calidad de un modelo de regresión logística, se analiza tanto el modelo en su conjunto como los predictores que lo forman.

En el resumen del modelo se encuentran los índices de ajuste, como residuos, las devianzas y el AIC, los cuales vamos a utilizar para evaluar el ajuste del modelo.

R calcula la devianza nula (solo con la constante) y la devianza residual (todo el modelo). Para que el modelo sea bueno la residual debe ser menor que la nula ya que valores más bajo de -2LL (-2 veces el logaritmo de la verosimilitud, devianza) indican que el modelo predice la variable respuesta con mayor precisión.

Entonces, se considera que el modelo es útil si es capaz de mostrar una mejora explicando las observaciones respecto al modelo nulo (sin predictores). El *test Likelihood ratio* calcula la significación de la diferencia de residuos entre el modelo de interés y el modelo nulo. El estadístico sigue una distribución chi cuadrado con grados de libertad equivalentes a la diferencia de grados de libertad de los dos modelos.

Para el contraste del modelo nulo sobre el modelo con las variables predictoras aplicamos el método basado en la *deviance* mediante la función `anova()`, a la que añadimos la opción `test`="Chi" para que nos proporcione el nivel crítico de acuerdo con la distribución chi cuadrado.

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: ACTIVOS
```

```
Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                   74861      59705
SEXO    1      226     74860      59479   <2e-16 ***
EDAD   10     1852     74850      57627   <2e-16 ***
ECIVIL  3      316     74847      57310   <2e-16 ***
ESTU    6     2061     74841      55249   <2e-16 ***
NAC     2       77     74839      55173   <2e-16 ***
CCAA_  18     1288     74821      53884   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con esta función podemos observar la significatividad de cada término añadido.

La reducción de la *deviance* a 53,884 es importante (p-valor=2e-16 < 0.05) ya que el nivel crítico es muy pequeño. Entonces, podemos rechazar la hipótesis de nulidad para explicar la probabilidad de éxito (parados), se acepta el modelo en el estado actual sobre el nulo.

Acorde a los p-valores resultantes, todas nuestras variables son estadísticamente significativas. También son significativas las contribuciones al modelo de las categorías, exceptuando algunas categorías de la variable `Comunidad autónoma` como: Cantabria, Castilla y León, Cataluña, País Vasco y la Rioja, lo que sugiere que los cambios en estas categorías no están asociados con cambios en la respuesta.

En general, podemos decir que el modelo tiene un aporte significativo en el análisis.

En este capítulo utilizamos referencias como : Agresti (2007), Crawley (2007), Bendixen (2003).

Capítulo 5

Resultados: Datos EES

En este capítulo se analiza la Encuesta cuatrienal de Estructura Salarial de 2014. En la primera parte se realiza un análisis descriptivo y exploratorio de los datos donde comparamos la `Ganancia por hora` de los individuos según el `Sexo` y en función de otras variables socioeconómicas y características del empleo. Por último, se aplica inferencia para conocer y cuantificar los factores que influyen en el salario y la brecha salarial entre hombres y mujeres.

5.1. Análisis exploratorio de los datos

Salario por hora

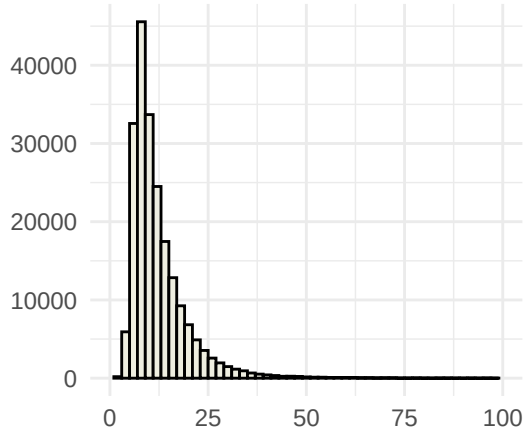


Figura 5.1. Distribución salario por hora.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2     7.6    10.1    12.5    14.6   699.0
```

La `Ganancia media por hora` de los trabajadores en España es de 12.46€. En la Figura 5.1 se observa que la variable tiene una distribución asimétrica hacia la derecha, donde el 75 % de los trabajadores ganan 14.62€ o menos.

Salario por hora en función de la variable sexo

En la siguiente salida tenemos un descriptivo de la variable `Salario por hora` respecto del `Sexo`, observando una diferencia en medias de 2.5€.

```
$hombre
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.2     8.2    10.9    13.5    15.8   699.0

$mujer
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.28    6.95    9.10   11.03   13.02  209.02
```

Aplicamos un boxplot para observar la distribución de cada uno de los grupos. Parece existir una diferencia en el salario entre hombres y mujeres.

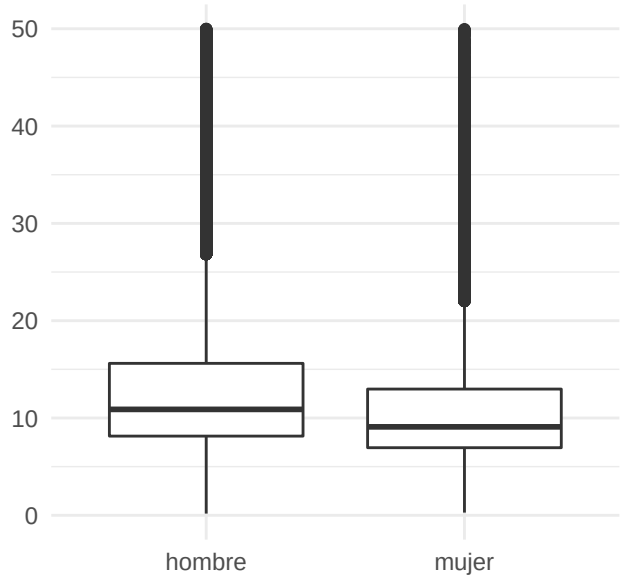


Figura 5.2. Salarios por hora en función del sexo.

Salario medio por hora según el sexo y la edad

En la Figura 5.3 observamos que desde el principio existe una pequeña diferencia entre hombres y mujeres en la ganancia por hora para aquellos con menos de 19 años (7.43€ vs 7.22€ por hora, diferencia de 0.21€). Esa diferencia se va haciendo mayor según avanza en grupos de `Edad`, siendo la diferencia de 4.81€ en individuos con más de 59 años.

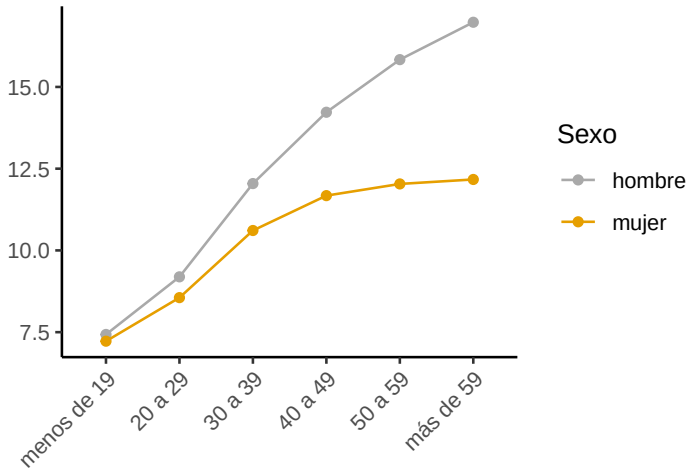


Figura 5.3. Salarios por hora en función del sexo y edad.

Salario por hora según el sexo y el estudio

Esta variable se clasifica de la siguiente manera:

Cuadro 5.1: Niveles de estudio.

Cód.	Descripción
1	Menos que primaria
2	Educación primaria
3	Primera etapa de educación secundaria
4	Segunda etapa de educación secundaria
5	Enseñanzas de formación profesional de grado superior y similares
6	Diplomados universitarios y similares
7	Licenciados y similares, y doctores universitarios

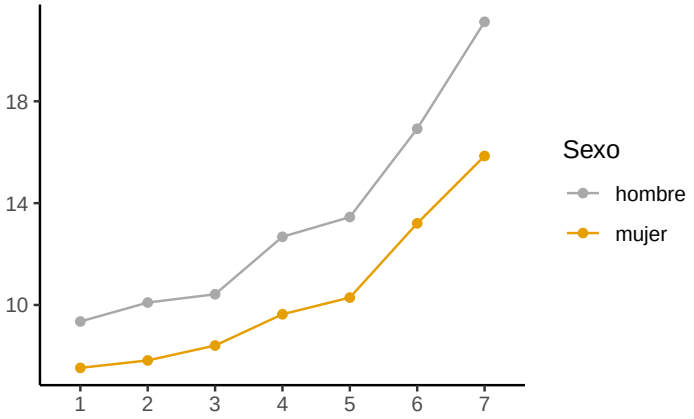


Figura 5.4. Salarios por hora en función del sexo y estudio.

Se observa en la Figura 5.4 que ambas líneas de tendencia aumentan, aun así el `Salario por hora` de los hombres crece ligeramente más rápido al pasar a los niveles de mayor formación. La brecha salarial parece presentar cambios significativos en los niveles de formación más elevados.

Salario medio por hora según el sexo y el año de antigüedad

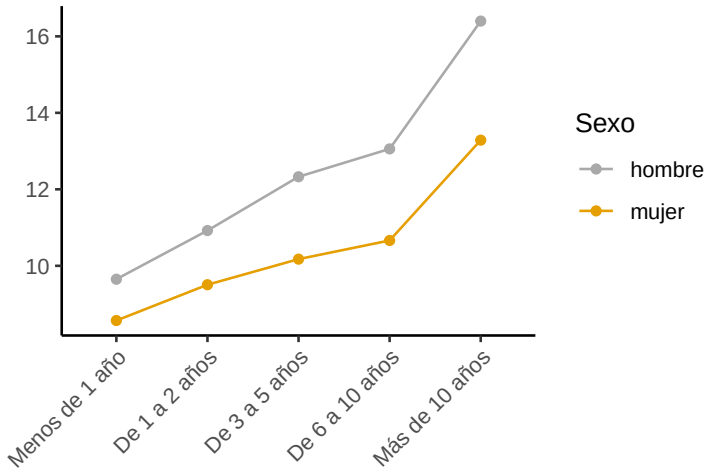


Figura 5.5. Salarios por hora en función del sexo y anti.

A mayor `antigüedad` aumenta el `Salario medio por hora`. En cuanto a la brecha del salario entre hombre y mujeres, se ensancha ligeramente según avanza el tiempo de estancia en la empresa.

Salario medio por hora según el sexo y el tipo jornada laboral

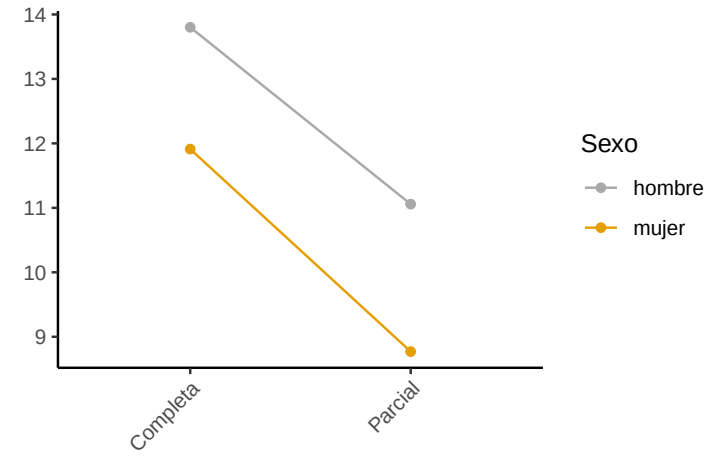


Figura 5.6. Salarios por hora en función del sexo y jornada.

La tendencia del `Salario medio por hora` tanto en hombres como en mujeres disminuye al pasar de tener jornada de tiempo completo a jornada de tiempo parcial, aumentando ligeramente la diferencia, aun así no parece ser significativa.

Salario medio por hora según el sexo y el tipo contrato

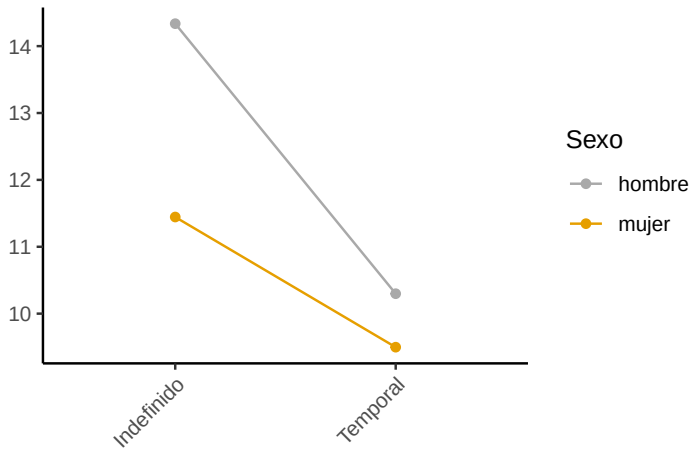


Figura 5.7. Salarios por hora en función del sexo y contrato.

Se observa en la Figura 5.7 que el salario disminuye al pasar de tener un contrato indefinido a uno determinado, asimismo, disminuye la brecha salarial, pasando de 2.86€ a 0.8€. En consecuencia, podríamos decir que las mujeres con contrato temporal estarían menos discriminadas con respecto a los hombres con contrato temporal que las mujeres que tienen contrato indefinido con respecto a los hombres de esa categoría.

Salario medio por hora según el sexo y si tiene o no responsabilidad en el trabajo

El tener `Responsabilidades de supervisión` frente a no tenerlas aumenta la brecha salarial entre hombres y mujeres. La brecha salarial pasa de 1.62€ a 4.18€.

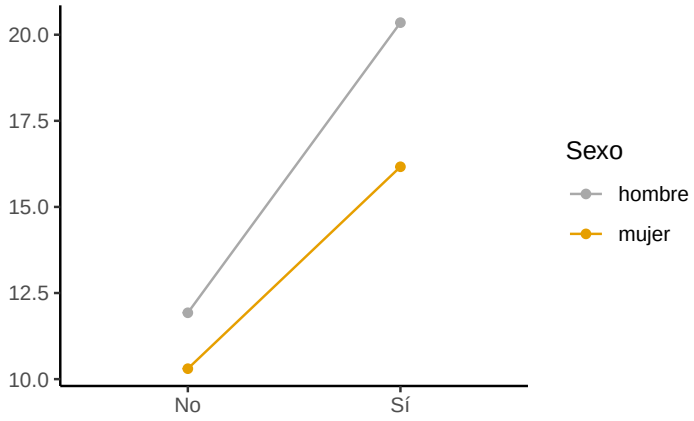


Figura 5.8. Salarios por hora en función del sexo y respon.

Salario medio por hora según el sexo y la ocupación

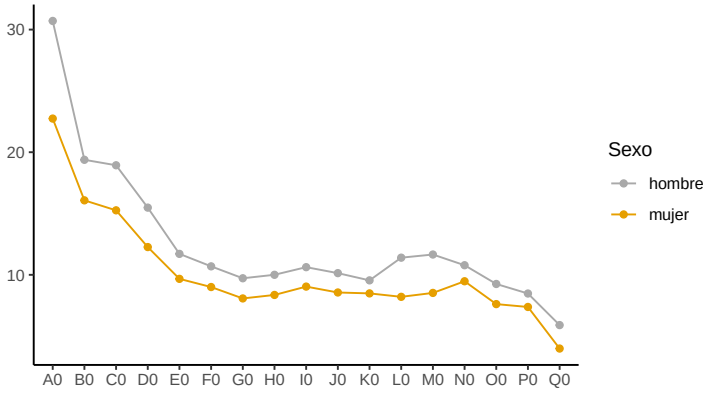


Figura 5.9. Salarios por hora en función del sexo y ocupación.

En la Figura 5.9 se visualiza que las brechas más importantes del `Salario medio por hora` entre hombres y mujeres de las distintas `ocupaciones` se encuentra entre las ocupaciones de: "Directores y gerentes", "Técnicos y profesionales científicos e intelectuales", "Trabajadores cualificados de las industrias manufactureras" y "operadores de instalaciones y máquinas".

5.2. Inferencia

5.2.1. Comparación de medias: prueba de Wilcoxon

Se utiliza un contraste de hipótesis para comparar las distribuciones del `Salario por hora` respecto a la variable `Sexo`. Queremos contrastar si la retribución es igual entre hombres y mujeres o si existen diferencias.

Tras visualizar los gráficos y verificar la asimetría de las distribuciones se aplica el *test* no paramétrico para dos muestras, el *test de Wilcoxon* para muestras independientes.

Sea $X_1, X_2, , X_n$ una muestra de tamaño n procedente de una población con distribución de tipo continuo F_X y sea $Y_1, Y_2, , Y_m$ una muestra de tamaño m de otra población con distribución de tipo continuo F_Y, independiente de la anterior.

Se trata de contrastar la hipótesis nula $H_0 : F_X = F_Y$ (las muestras proceden de la misma distribución con respecto a la mediana) frente a la alternativa bilateral $H_1 : F_X \neq F_Y$ o bien, frente a alternativas unilaterales de diferencias en la posición central o mediana.

El *test de Wilcoxon* se basa en los rangos o posiciones de las observaciones ordenadas de menor a mayor. Utiliza la idea de que, si H_0 es cierta, se espera que los rangos correspondientes a los valores de una y otra muestra, ordenados conjuntamente, estén entremezclados o dispersos, mientras que en otro caso, debe esperarse que los rangos de las observaciones de cada muestra estén muy agrupados en los extremos. Por ejemplo, si $Me_X < Me_Y$ la suma de los rangos de X tenderá a ser pequeña y si $Me_x > Me_y$ la suma de los rangos de X tenderá a ser grande.

Llamando R a la suma de los rangos asociados a las observaciones de una cualquiera de las dos muestras (por ejemplo, la de menor tamaño, que supongamos que es X), se define la suma de los rangos de los elementos de la muestra X:

$$R_X = \sum_{j=1}^{n+m} jI_j,$$

siendo I_j= 1{la observación con rango j proviene de la muestra X}.

En R el *test de Wilcoxon* se obtiene con la función:

```
wilcox.test(data_ees$SAL ~ data_ees$SEXO)
```

```
    Wilcoxon rank sum test with continuity correction

data:  data_ees$SAL by data_ees$SEXO
W = 6.5e+09, p-value <2e-16
alternative hypothesis: true location shift is not equal to 0
```

Se obtiene un p-valor menor al nivel de significación del 5 %, por lo tanto hay evidencia muestral suficiente para rechazar la hipótesis nula y aceptar la alternativa. Se rechaza que los salarios por hora son iguales en ambos sexos.

5.2.2. Modelos de regresión para estimar la diferencia y la discriminación salarial

La estimación de la diferencia y discriminación salarial se basa en la especificación, estimación e interpretación de modelos de regresión lineales, con los que se puede explicar cómo incide cada característica de la persona empleada en el salario y, en concreto, cómo afecta si el trabajador es hombre o mujer al salario percibido. Es decir, interesa conocer cuál es la discriminación salarial de las mujeres, para las diferentes características (personales y laborales) de los trabajadores.

Se lleva a cabo dos pasos para el análisis:

1) La estimación del efecto de las diferentes variables sobre el `Salario por hora`. En especial el efecto del `Sexo`, para comprobar si el ser hombre o mujer condiciona el salario que percibe y, en consecuencia, se estima la discriminación salarial media por hora de las mujeres en el mercado de trabajo.

2) La estimación de las interacciones entre la variable `Sexo` y el resto de características de los trabajadores. Esto permite matizar (al alza o a la baja) la estimación de la discriminación salarial media de las mujeres, obtenida con el modelo anterior, determinando qué características inciden (más o menos) en la menor retribución salarial de las mujeres.

Variables del modelo

- Variable dependiente: la variable a explicar es el `Salario por hora`, aproximado por la variable `Ganancia media por hora`.

  Se aplica la transformación logarítmica con el objetivo de eliminar el efecto de las unidades de medida y linealizar la relación entre las variables, de manera que aumente la bondad del ajuste.

- Variables independientes: el logaritmo neperiano de la variable dependiente queda explicado por las variables categóricas verificadas en el Cuadro 5.2 y la variable `Ocupación`:

Cuadro 5.2: Variables explicativas

Variable	Cód	Descripción
Sexo	0	Hombre
	1	Mujer
Grupos de edad	1	menos de 19
	2	20 a 29
	3	30 a 39
	4	40 a 49
	5	50 a 59
	6	más de 59
Antiguedad	1	Menos de 1 año
	2	1 a 2 años
	3	3 a 5 años
	4	6 a 10 años
	5	más de 10 años
Tipo de jornada	1	Completo
	2	Parcial
Tipo de contrato	1	Indefinidad
	2	Determinada
Responsabilidad de supervisión	0	Sí
	1	No

Note:
Para el Nivel de estudio : Ver Cuadro 5.1

Cód.	Descripción **variable ocupación**
A0	Directores y gerentes
B0	Técnicos y profesionales científicos e intelectuales de la salud y la enseñanza
C0	Otros técnicos y profesionales científicos e intelectuales
D0	Técnicos; profesionales de apoyo
E0	Empleados de oficina que no atienden al público
F0	Empleados de oficina que atienden al público
G0	Trabajadores de los servicios de restauración y comercio
H0	Trabajadores de los servicios de salud y el cuidado de personas
I0	Trabajadores de los servicios de protección y seguridad
J0	Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero
K0	Trabajadores cualificados de la construcción, excepto los operadores de máquinas
L0	Trabajadores cualificados de las industrias manufactureras, excepto operadores de instalaciones y máquinas
M0	Operadores de instalaciones y maquinaria fijas, y montadores
N0	Conductores y operadores de maquinaria móvil
O0	Trabajadores no cualificados en servicios
P0	Peones de la agricultura, pesca, construcción, industrias manufactureras y transportes
Q0	Ocupaciones militares

Utilizamos 8 variables categóricas y se estiman un total de 35 coeficientes.

5.2.2.1. Modelo 1

Se plantea una regresión lineal para explicar cómo influye cada variable en la formación de los salarios percibidos. Qué características personales y laborales influyen en la determinación de la estructura salarial, en qué sentido y en qué cuantía actúa cada una de ellas.

Este modelo proporciona, una estimación de la discriminación salarial media por hora en el puesto de trabajo por razón de sexo del mercado laboral español, es decir, qué parte de la diferencia salarial entre mujeres y hombres se debe únicamente al sexo (parte discriminatoria) y no está ligada a otro tipo de características.

Interpretación de los coeficientes

Considerando que todas las variables son discretas, los coeficientes de las categorías de cada variable explicativa que quedan en el modelo estimable se interpretan como el aumento o disminución porcentual que experimenta la `Ganancia media por hora`, según sea el signo positivo o negativo, respectivamente, que tienen las categorías que quedan en el modelo con relación a la categoría de referencia.

Por lo tanto, a la hora de estimar el modelo siempre existe un individuo de referencia, cuyas características vienen dadas por las categorías que han quedado fuera. En nuestro caso, la función que utilizaremos en R asume como categoría de referencia la primera de la variable, generalmente por orden alfabético.

Estimación del modelo y resultados

Una vez que fijadas las categorías de referencias el logaritmo neperiano de la `Ganancia media por hora` queda explicado por 8 variables, estimándose 35 coeficientes cuya interpretación se realiza con relación a la categoría de referencia.

Las características de referencia son las siguientes:

Características personales:

- Hombre.
- Menor de 19 años de edad.
- Estudios de menos que primaria

Características laborales:

- Con menos de un año de antigüedad en la empresa.
- Trabajadores no cualificados en servicios.[1]
- Con jornada completa.
- Con contrato indefinido.
- Sin responsabilidad en la organización y/o supervisión de otras personas.

Por ejemplo, en el caso de la variable `Sexo`, la categoría de referencia es hombre, el coeficiente que acompaña a la categoría `Sexo:mujer` indica el efecto de más o de menos sobre el salario que tiene ser mujer con relación a ser hombre.

En la siguiente salida se presentan los coeficientes resultantes de la estimación del modelo ajustado.

```
modelo1 <- lm(log(SAL) ~ SEXO + EDAD + ESTU + ANTI + TIPOJOR + TIPOCON +
    RESPONSA + OCUP)
summary(modelo1)
```

[1] Es el único caso en el cual cambiamos el grupo de referencia para facilitar la interpretación. Este grupo tienen un gran número de individuos, lo que garantiza que sea una referencia que se identifique con la generalidad de los trabajadores, siendo mejor para la comparación.

```
Call:
lm(formula = log(SAL) ~ SEXO + EDAD + ESTU + ANTI + TIPOJOR +
    TIPOCON + RESPONSA + OCUP)

Residuals:
   Min     1Q Median     3Q    Max
-3.571 -0.230 -0.021  0.207  4.155

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.88378    0.02278   82.70  < 2e-16 ***
SEXO1       -0.15916    0.00183  -86.83  < 2e-16 ***
EDAD2        0.00962    0.02153    0.45  0.65493
EDAD3        0.08235    0.02149    3.83  0.00013 ***
EDAD4        0.14465    0.02151    6.72  1.8e-11 ***
EDAD5        0.20754    0.02155    9.63  < 2e-16 ***
EDAD6        0.24607    0.02178   11.30  < 2e-16 ***
ESTU2        0.00533    0.00724    0.74  0.46163
ESTU3        0.03951    0.00714    5.53  3.1e-08 ***
ESTU4        0.12917    0.00725   17.83  < 2e-16 ***
ESTU5        0.17657    0.00757   23.32  < 2e-16 ***
ESTU6        0.22902    0.00775   29.55  < 2e-16 ***
ESTU7        0.35140    0.00764   45.97  < 2e-16 ***
ANTI2        0.01524    0.00319    4.77  1.8e-06 ***
ANTI3        0.06813    0.00336   20.25  < 2e-16 ***
ANTI4        0.12180    0.00317   38.42  < 2e-16 ***
ANTI5        0.25244    0.00322   78.47  < 2e-16 ***
TIPOJOR2    -0.07780    0.00232  -33.55  < 2e-16 ***
TIPOCON2    -0.03445    0.00244  -14.12  < 2e-16 ***
RESPONSA1    0.15036    0.00249   60.41  < 2e-16 ***
OCUPA0       0.58926    0.00633   93.03  < 2e-16 ***
OCUPB0       0.42396    0.00531   79.83  < 2e-16 ***
OCUPC0       0.32818    0.00493   66.60  < 2e-16 ***
OCUPD0       0.22244    0.00414   53.79  < 2e-16 ***
OCUPE0       0.05841    0.00438   13.33  < 2e-16 ***
OCUPF0       0.04625    0.00506    9.13  < 2e-16 ***
OCUPG0       0.02768    0.00455    6.08  1.2e-09 ***
OCUPH0       0.00706    0.00482    1.46  0.14314
OCUPI0       0.04280    0.00616    6.95  3.7e-12 ***
OCUPJ0       0.01436    0.01400    1.03  0.30507
OCUPK0       0.04923    0.00573    8.59  < 2e-16 ***
OCUPL0       0.10294    0.00437   23.54  < 2e-16 ***
OCUPM0       0.12337    0.00456   27.08  < 2e-16 ***
OCUPN0       0.11257    0.00515   21.84  < 2e-16 ***
OCUPP0      -0.01355    0.00514   -2.63  0.00841 **
OCUPQ0      -0.43586    0.04257  -10.24  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.363 on 209400 degrees of freedom
Multiple R-squared:  0.463,   Adjusted R-squared:  0.463
F-statistic: 5.16e+03 on 35 and 209400 DF,  p-value: <2e-16
```

Se observa que la brecha salarial del modelo ajustado es de -15.91 %.

El valor y signo de los coeficientes del modelo confirman los resultados obtenidos en el análisis exploratorio sobre el comportamiento de la tendencia, al alza o a la baja, del salario por hora en función de las categorías de cada una de las variables del análisis con respecto a la categoría de referencia.

El modelo estimado explica el 46.3 % de la variabilidad de la variable endógena y sus coeficientes son representativos excepto el `Grupo de edad` de 20 a 29 años, el `Grupo de estudio` "Educación primaria", las `Ocupaciones` "Trabajadores de los servicios de salud y el cuidado de personas" y "Trabajadores cualificados en el sector agrícola, ganadero, forestal y pesquero". El `Salario por hora` percibido por los individuos de estas categorías y sus categorías de referencia son similares.

5.2.2.2. Modelo 2

Una vez se verifica la existencia de discriminación salarial en función del `Sexo` en el mercado de trabajo español, así como cuantificado el efecto de las diferentes variables sobre el salario, interesa medir la discriminación salarial indirecta que resulta de la interacción del `Sexo` del individuo con otras características.

En concreto, se estima la discriminación salarial por razón de `Sexo` desagregada en función de las características personales, laborales que la provocan. Con ello, se intenta determinar las causas que debilitan o intensifican la brecha salarial entre mujeres y hombres, que no está justificada por su diferente dotación de características.

Interpretación de los coeficientes

Los coeficientes de la primera parte (β_1 a β_{35}) se interpretan de la misma manera que los del Modelo 1, indican el aumento o disminución porcentual que experimenta el salario de la categoría en cuestión respecto al de la categoría que ha quedado fuera del modelo. Pero en el Modelo 2 la información de estos coeficientes se completa con la información que proporcionan los β_{36} a β_{69}. Estos últimos también incorporan una parte del efecto de las variables de la primera parte.

Los coeficientes de la segunda parte de la ecuación (β_{36} a β_{69}) se interpretan de forma semejante, teniendo en cuenta dentro de cada variable cruzada las categorías que se han quedado fuera, que son los hombres de la categoría en cuestión y los hombres y las mujeres de la categoría que ha quedado fuera en la primera parte del modelo. Estos coeficientes cuantifican diferencias salariales entre hombres y mujeres idénticos con las mismas características personales y laborales.

Por ejemplo, el coeficiente que acompaña a la categoría "mujeres con contrato temporal", se interpreta en relación a las categorías eliminadas (de referencia) de la variable `Tipo de contrato cruzada con Sexo`, es decir, en relación a hombres con contrato temporal y hombres y mujeres con contrato indefinido. De esta forma, este coeficiente indica si el aumento o disminución porcentual que experimenta el salario de las mujeres con contrato temporal sobre los hombres con contrato temporal es diferente al salario de las mujeres con contrato indefinido sobre los hombres con contrato indefinido.

De este modo, el coeficiente que acompaña a esta categoría mide cuánto aumenta o disminuye la parte del salario que depende del sexo en las mujeres con contrato temporal con relación a la parte del salario explicada por el sexo de las mujeres con contrato indefinido, ya que el coeficiente aísla el efecto del resto de variables.

En concreto, este coeficiente ofrece información acerca de si las mujeres ocupadas con contrato temporal están más o menos discriminadas que las mujeres que tienen contrato indefinido. Si el coeficiente fuera de signo negativo, las mujeres con contrato temporal estarían más discriminadas que las mujeres con contrato indefinido, mientras que ocurriría lo contrario si el signo fuera positivo, indicando el valor del coeficiente cuánto más o menos discriminadas estarían.

Estimación del modelo y resultados

En la siguiente salida se muestran los coeficientes estimados por el Modelo 2 para cada categoría de las variables. Los correspondientes a la primera parte de la ecuación se interpretan de manera similar a los

del Modelo 1. De hecho, los coeficientes estimados toman un valor muy similar, derivándose las mismas conclusiones del análisis de ambos.

```
modelo2 <- lm(log(SAL) ~ SEXO * EDAD + SEXO * ESTU + SEXO * ANTI +
    SEXO * TIPOJOR + SEXO * TIPOCON + SEXO * RESPONSA + SEXO * OCUP)
summary(modelo2)
```

```
Call:
lm(formula = log(SAL) ~ SEXO * EDAD + SEXO * ESTU + SEXO * ANTI +
    SEXO * TIPOJOR + SEXO * TIPOCON + SEXO * RESPONSA + SEXO *
    OCUP)

Residuals:
   Min     1Q Median     3Q    Max
-3.549 -0.228 -0.021  0.206  4.187

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.823524   0.029165   62.52  < 2e-16 ***
SEXO1           0.020170   0.046711    0.43  0.66589
EDAD2           0.056426   0.027540    2.05  0.04048 *
EDAD3           0.139300   0.027490    5.07  4.0e-07 ***
EDAD4           0.202124   0.027512    7.35  2.0e-13 ***
EDAD5           0.271283   0.027562    9.84  < 2e-16 ***
EDAD6           0.322684   0.027822   11.60  < 2e-16 ***
ESTU2           0.008239   0.008749    0.94  0.34639
ESTU3           0.037534   0.008623    4.35  1.3e-05 ***
ESTU4           0.143987   0.008763   16.43  < 2e-16 ***
ESTU5           0.197761   0.009140   21.64  < 2e-16 ***
ESTU6           0.259938   0.009601   27.07  < 2e-16 ***
ESTU7           0.377952   0.009421   40.12  < 2e-16 ***
ANTI2           0.010445   0.004193    2.49  0.01273 *
ANTI3           0.069681   0.004489   15.52  < 2e-16 ***
ANTI4           0.131304   0.004239   30.98  < 2e-16 ***
ANTI5           0.262774   0.004199   62.58  < 2e-16 ***
TIPOJOR2       -0.091387   0.003787  -24.13  < 2e-16 ***
TIPOCON2       -0.044909   0.003330  -13.49  < 2e-16 ***
RESPONSA1       0.165419   0.003123   52.97  < 2e-16 ***
OCUPA0          0.571565   0.008723   65.52  < 2e-16 ***
OCUPB0          0.351940   0.008780   40.09  < 2e-16 ***
OCUPC0          0.293227   0.007353   39.88  < 2e-16 ***
OCUPD0          0.200412   0.006355   31.53  < 2e-16 ***
OCUPE0          0.025619   0.007066    3.63  0.00029 ***
OCUPF0          0.003921   0.008658    0.45  0.65066
OCUPG0         -0.009006   0.007744   -1.16  0.24481
OCUPH0         -0.032964   0.009112   -3.62  0.00030 ***
OCUPI0          0.027018   0.007811    3.46  0.00054 ***
OCUPJ0          0.000442   0.015528    0.03  0.97729
OCUPK0          0.046059   0.007174    6.42  1.4e-10 ***
OCUPL0          0.103085   0.006235   16.53  < 2e-16 ***
OCUPM0          0.142484   0.006556   21.73  < 2e-16 ***
OCUPN0          0.104710   0.006737   15.54  < 2e-16 ***
OCUPP0         -0.012303   0.007075   -1.74  0.08207 .
```

```
OCUPQ0           -0.411025   0.043700   -9.41  < 2e-16 ***
SEXO1:EDAD2      -0.133382   0.043968   -3.03  0.00242 **
SEXO1:EDAD3      -0.151898   0.043896   -3.46  0.00054 ***
SEXO1:EDAD4      -0.153807   0.043929   -3.50  0.00046 ***
SEXO1:EDAD5      -0.169032   0.044016   -3.84  0.00012 ***
SEXO1:EDAD6      -0.200926   0.044517   -4.51  6.4e-06 ***
SEXO1:ESTU2      -0.012282   0.015497   -0.79  0.42804
SEXO1:ESTU3      -0.005107   0.015292   -0.33  0.73839
SEXO1:ESTU4      -0.056651   0.015520   -3.65  0.00026 ***
SEXO1:ESTU5      -0.086708   0.016241   -5.34  9.4e-08 ***
SEXO1:ESTU6      -0.092530   0.016456   -5.62  1.9e-08 ***
SEXO1:ESTU7      -0.078027   0.016262   -4.80  1.6e-06 ***
SEXO1:ANTI2       0.008641   0.006448    1.34  0.18019
SEXO1:ANTI3      -0.008706   0.006765   -1.29  0.19811
SEXO1:ANTI4      -0.025487   0.006381   -3.99  6.5e-05 ***
SEXO1:ANTI5      -0.030563   0.006522   -4.69  2.8e-06 ***
SEXO1:TIPOJOR2    0.015854   0.004813    3.29  0.00099 ***
SEXO1:TIPOCON2    0.023294   0.004894    4.76  1.9e-06 ***
SEXO1:RESPONSA1 -0.042877   0.005174   -8.29  < 2e-16 ***
SEXO1:OCUPA0      0.029221   0.013254    2.20  0.02748 *
SEXO1:OCUPB0      0.136864   0.011211   12.21  < 2e-16 ***
SEXO1:OCUPC0      0.083401   0.010101    8.26  < 2e-16 ***
SEXO1:OCUPD0      0.057021   0.008590    6.64  3.2e-11 ***
SEXO1:OCUPE0      0.072987   0.009130    7.99  1.3e-15 ***
SEXO1:OCUPF0      0.079801   0.010770    7.41  1.3e-13 ***
SEXO1:OCUPG0      0.062473   0.009632    6.49  8.8e-11 ***
SEXO1:OCUPH0      0.068142   0.010831    6.29  3.2e-10 ***
SEXO1:OCUPI0      0.069889   0.015974    4.38  1.2e-05 ***
SEXO1:OCUPJ0      0.064573   0.040803    1.58  0.11352
SEXO1:OCUPK0      0.032752   0.033503    0.98  0.32828
SEXO1:OCUPL0     -0.086005   0.010822   -7.95  1.9e-15 ***
SEXO1:OCUPM0     -0.093432   0.009807   -9.53  < 2e-16 ***
SEXO1:OCUPN0      0.075750   0.019967    3.79  0.00015 ***
SEXO1:OCUPP0      0.002540   0.011367    0.22  0.82321
SEXO1:OCUPQ0     -0.121353   0.213322   -0.57  0.56944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.361 on 209366 degrees of freedom
Multiple R-squared:  0.467,   Adjusted R-squared:  0.466
F-statistic: 2.65e+03 on 69 and 209366 DF,  p-value: <2e-16
```

La importancia del `Sexo` en la determinación de los salarios es clara. Sin embargo, ciertas características personales o concernientes al puesto de trabajo aumentan o reducen esta importancia. En determinados perfiles de trabajadores, el salario es más dependiente del sexo que en otros, por lo que hay mujeres que por sus características personales y laborales, están más discriminadas que otras.

Al analizar el cruce de la variable `Sexo y Edad`, obtenemos coeficientes negativos en todas sus categorías, con valores cada vez más grande. Lo que significa que al pasar de los grupos de referencia de este cruce de variables (hombres en el grupo de edad de la categoría de estudio y hombres y mujeres en el grupo de menos de 19 años) a las distintas categorías de estudio, cada vez, las mujeres están más discriminadas. Esto manteniendo lo demás constante. Confirmando lo visualizado en la Figura 5.3, donde se observa que la discriminación va aumentando progresivamente conforme aumenta la `Edad`.

Con respecto a la variable `Nivel de estudios` cruzada con `Sexo`, se interpreta en relación a hombres en la

categoría de estudio en cuestión y hombres y mujeres con nivel de Menos que primaria. De acuerdo a los resultados las mujeres con estudios más avanzados estarían más discriminadas que las mujeres con estudios de menos que primaria.

Con el aumento del nivel de estudio aumenta la discriminación salarial, siendo importante destacar que los dos grupos siguientes a la categoría de referencia (Educación primaria y Primera etapa de educación secundaria) son no significativos, lo que quiere decir que la mujer soporta el mismo nivel de discriminación salarial en cualquiera de estas categorías.

En el caso de la `Antigüedad` en la empresa perjudica a la mujer trabajadora, aumentándose la brecha salarial entre hombres y mujeres con las mismas características de forma progresiva conforme se eleva el número de años de antigüedad.

De acuerdo a los resultados esto sucede especialmente a partir de los 6 años de estancia, ya que en los dos primero grupos sus p-valores están por encima del 5 % de significación. Por lo que las mujeres en estos grupos se encuentran igual de discriminadas.

De acuerdo al `Tipo de jornada`, las mujeres con jornada parcial estarían menos discriminadas que las mujeres con jornada de tiempo completo.

Las mujeres contratadas por tiempo indefinido están relativamente peor tratadas respecto al hombre que las trabajadoras con contrato de duración determinada. La ganancia media por hora de las mujeres con contrato parcial aumenta en 2.3 % con respecto a las variables de referencia.

Las mujeres con `Responsabilidad de supervisión` de otras personas están más discriminadas que las que no las asumen.

En cuanto al cruce de `Sexo y Ocupación` el grupo de referencia serían los hombres de la ocupación en cuestión y los hombres y las mujeres que son trabajadores no cualificados en servicio. En casi todos los grupos de estudios las mujeres son mejores tratadas en comparación con los hombres de esas categorías estudiadas y las mujeres y hombre trabajadores no cualificados en servicio. Excepto para las categorías de “Trabajadores cualificados de las industrias manufactureras, operadores de instalaciones y máquinas y Operadores de instalaciones y maquinaria fijas, y montadores”.

Al examinar los coeficientes del modelo se observa que la categorías no representativas se encuentran en los grupos 2 y 3 de la variable de `Estudios`, las dos primeras categorías de `Antigüedad` y en las categorías de `ocupación` como JO, KO, PO, y QO.

El coeficiente R^2 nos dice que con las variables podemos explicar el 46.7 % de la variabilidad en el cambio del salario medio por hora al aplicar cruces de variables personales y laborales con el sexo.

Se realiza el diagnóstico de los supuestos de normalidad y homogeneidad utilizando métodos gráficos. Ambos se cumplen de forma aproximada. El modelo no presenta un patrón específico de variabilidad cambiante al graficar los residuos frente a los valores ajustados. En el gráfico de cuantiles la parte central de la curva se ajusta razonablemente a la diagonal.

En este capítulo se utilizan las referencia CEOE (2019), Anghel et al. (2018), De Cabo y Rodríguez (2014), De Cabo et al. (2013), De Cabo y Garzón (2007), Crawley (2007).

Capítulo 6

Minería de datos

La clasificación o aprendizaje supervisado es la tarea de asignar objetos a una de varias categorías predefinidas a partir de una muestra de aprendizaje. Es un problema generalizado que abarca cualquier aplicación diversa, los ejemplos incluyen: la detección de mensajes de correo electrónico no deseado basados en el encabezado y el contenido del mensaje, la clasificación de las células como malignas o benignas en función de los resultados de las imágenes de resonancia magnética y la clasificación de galaxias en función de sus formas.

Cada uno de los registros en el conjunto de entrenamiento consiste en un vector de entrada representado por x_i y una etiqueta de salida representada por y_i. El vector x_i en un conjunto de características representativas del punto i de la muestra. La tarea del algoritmo de clasificación es inferir una función f (de un posible conjunto dado de funciones F) que puede asignar los x_i a los respectivos y_i, con un alto nivel de precisión. El proceso de inferir f, usando los datos de entrenamiento se llama aprendizaje. Una vez que el modelo está capacitado, utilizamos este modelo aprendido con los nuevos registros para identificar nuevas etiquetas. La capacidad de tal modelo para identificar correctamente las etiquetas del nuevo conjunto, que difieren del conjunto de entrenamiento, se conoce como eficiencia de la predicción.

Se distinguen tres fases a la hora de realizar un proceso de clasificación: la definición de las clases, la representación de la información mediante atributos y el aprendizaje mediante algoritmos.

En este capítulo se aplican técnicas de clasificación supervisada utilizando la variable `Activos`, para identificar si el individuo perteneciente a la población económicamente activa se encuentra sin trabajo, a partir de las variables estudiadas, y averiguaremos cuál es la importancia de los distintos predictores.

Se obtiene un resultado binario, representado como "0" si está ocupado y "1" si se encuentra parado. Siendo importante destacar que los grupos se encuentran desbalanceados. El 86.34 % de los individuos se encuentran ocupados.

```
ocupados  parados
  0.8634   0.1366
```

Al final del capítulo se comparan los resultados obtenidos de cada uno de los modelos analizados.

6.1. Conjuntos de entrenamiento y prueba

El conjunto de entrenamiento se utiliza para construir un modelo de clasificación. Consiste en registros con etiquetas de clase desconocidas que posteriormente se aplica al conjunto de prueba.

El subconjunto de datos de entrenamiento es utilizado para estimar los parámetros del modelo. Por su parte, el subconjunto de datos de prueba se emplea para comprobar el comportamiento del modelo estimado y evaluar la eficacia para hacer predicciones correctas.

Con nuestros datos preparados para el análisis, se define un conjunto de entrenamiento, en este caso el 80 % de los datos para enseñarle al computador como clasificar y el 20 % para evaluación. Cada registro de la base de datos debe aparecer en uno de los dos subconjuntos, y para dividir los datos se utiliza un procedimiento de muestreo. Se emplea la función `sample.split()` de la librería `caTools`. Fijamos semilla para obtener la misma muestra cada vez que arrojemos la función.

La orden `set.seed()` se utiliza para fijar el arranque de la secuencia aleatoria, de forma que si repetimos la orden con el mismo número volveremos a obtener exactamente la misma muestra, lo que permite reproducir los resultados exactos. Sin embargo pueden obtenerse muestras y resultados distintos con equipos, sistemas operativos, o versiones de R diferentes, aunque los resultados serán en general similares.

```
set.seed(123)
split <- sample.split(data_activos_clasif$ACTIVOS, SplitRatio = 0.8)

activos_train <- subset(data_activos_clasif, split == TRUE)
activos_test <- subset(data_activos_clasif, split == FALSE)
```

Para la creación y evaluación de los modelos, que se utilizarán más adelante, se extraen las variables objetivo y respuestas de la data de entrenamiento y prueba.

```
x_train <- activos_train[, -which(colnames(activos_train) == "ACTIVOS")]
x_test <- activos_test[, -which(colnames(activos_test) == "ACTIVOS")]

y_train <- factor(activos_train$ACTIVOS)
y_test <- factor(activos_test$ACTIVOS)
```

6.2. Métricas para analizar y comparar los algoritmos

Para medir el desempeño de un clasificador se utilizarán las siguientes métricas:

Matriz de confusión

Una matriz de confusión es una herramienta visual usada en el aprendizaje supervisado que permite la evaluación del desempeño de un modelo de clasificación. Cada fila representa el número de predicciones de cada clase, mientras que cada columna representa la clasificación real. Uno de los principales beneficios de las matrices de confusión es que permite ver si el sistema se confunde entre clases.

Cuadro 6.1. Matriz de confusión.

	ACTUAL	
PREDICCION	Positivo(1)	Negativo(0)
Positivo(1)	Verdadero Positivo(VP)	Falso Positivo (FP)
Negativo(0)	Falso Negativo (FN)	Verdadero Negativo(VN)

La Tabla 6.1 representa la matriz de confusión para un problema de clasificación binaria. Los valores de la diagonal representan las predicciones correctas para ambas clases. Los demás son errores, por ejemplo, FN es el número de registros de la clase 1 que se predice incorrectamente como clase 0, siendo FP el número de

registros de la clase 0 que se predice incorrectamente como clase 1.

Exactitud (accuracy)

Aunque una matriz de confusión proporciona la información necesaria para determinar qué tan bien se desempeña un modelo de clasificación, resumir esta información con un solo número haría más conveniente comparar el desempeño de diferentes modelos.

La métrica *accuracy* es la proporción del número total de predicciones que fueron clasificadas correctamente respecto al total.

$$Ac = \frac{VP + VN}{VP + FP + FN + VN}$$

Sensitivity

Es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea definido como positivo. Corresponde a los casos positivos bien clasificados sobre el total de casos reales positivos.

$$Sensitivity = \frac{VP}{VP + FN}$$

Specificity

Es la probabilidad de clasificar correctamente a un individuo cuyo estado real sea clasificado como negativo. Los resultados negativos verdaderos se dividen por la suma de todos los resultados negativos.

$$Specificity = \frac{VN}{VN + FP}$$

Accuracy Balanceado

La medida de Exactitud ha sido muy usada para determinar la eficiencia de un clasificador pero esta puede no ser una medida adecuada en el desempeño del mismo cuando hay clases altamente desequilibradas. Por ejemplo, supongamos que de 1000 casos en una muestra, 995 pertenecen a clase A y 5 a la clase B. Si el sistema clasifica a todos como A el *accuracy* sería del 99.5 % aún cuando el clasificador se equivoque en todos los de la clase B. Para este caso es recomendado usar el *accuracy* balanceado. Este indicador es la media entre *sensitivity* y *specificity*.

$$BalancedAccuracy = \frac{(sensitivity + specificity)}{2}$$

Roc

La curva ROC (*Receiver Operating Characteristic* o Característica Operativa del Receptor), conocida como la representación de sensibilidad frente a $1 - especificidad$, es una representación gráfica del rendimiento del clasificador que muestra la distribución de las fracciones de verdaderos positivos y de falsos positivos.

Una clasificación totalmente aleatoria daría un punto a lo largo de la línea diagonal. En un *test* perfecto, la curva ROC recorre los bordes izquierdo y superior del gráfico.

El área bajo la curva (AUC), que se corresponde con el *accuracy* balanceado explicado antes, es un indicador de rendimiento para los modelos de clasificación. Un AUC de 0.5 significa que el modelo no es mejor que una suposición aleatoria. Los valores de AUC por debajo de 0.5 resultan peores que si fuesen aleatorios. Un modelo perfecto tendría un AUC igual a 1.

Como habíamos visto anteriormente, hay una gran cantidad de personas que pertenecen a la categoría de ocupados, esto presenta un problema ya que puede provocar un sesgo por parte de los clasificadores a esta categoría. Es por esto que al comparar la efectividad de los modelos, se evaluará mayormente el *accuracy* balanceado y la curva ROC.

6.3. Técnicas de Clasificación Automática de datos

6.3.1. Algoritmo Naive Bayes

Naive Bayes (Maron, 1961) es un método de clasificación basado en el teorema de Bayes (teoría elemental de probabilidad). Este algoritmo permite estimar la probabilidad de un suceso a partir de la probabilidad de que ocurra otro suceso, del cual depende el primero. Este teorema funciona utilizando la probabilidad condicionada o la probabilidad de que algo suceda teniendo conocimiento de lo que ha ocurrido previamente.

El algoritmo esta basado en el supuesto de independencia entre las variables predictoras, es decir, que las diferentes características consideradas como variables explicativas no están relacionadas entre sí, calculando las probabilidades condicionales de cada predictor por separado. En particular, *Naive Bayes* supone que todas las características del conjunto de datos son igualmente importantes e independientes. Este supuesto es más que discutible en la mayoría de las aplicaciones prácticas, razón por la que recibe el apelativo de ingenuo (*naive*).

El clasificador *Naive Bayes* se puede entrenar de forma eficiente en un entorno de aprendizaje supervisado. En general, solo se requiere una pequeña cantidad de datos de entrenamiento para estimar los parámetros necesarios (medias y varianzas de las variables) para la clasificación, y el supuesto de independencia hace innecesario determinar la matriz completa de covarianzas. Pueden utilizarse como variables explicativas indistintamente variables cuantitativas y cualitativas. Supone distribución normal para los predictores cuantitativos.

Utilizando el teorema de Bayes se puede expresar la probabilidad de que un caso pertenezca a una clase determinada C (a posteriori, conocidos los valores de las variables explicativas) en términos de las probabilidades a priori:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

donde,

C, es la clase contra la que evaluaremos los nuevos datos.

X, nuevos casos, queremos saber qué tan probable es que proceda de una clase C.

$P(C)$, es la probabilidad a priori o incondicionada de pertenecer a la clase C.

$P(X)$, es la evidencia, lo que sabemos por medio de eventos pasados sobre la ocurrencia de X.

$P(X|C)$, es la probabilidad de que dada una clase C, X sea un caso de ella. Para poder calcularlo vamos a representar al individuo con una serie de características: $x_1, x_2, x_3, \ldots, x_n$. Vamos a asumir dos supuestos muy importantes: que no importa el orden de las características y que las probabilidades de cada característica dada una clase $C : P(x_i|c_j)$ son independientes entre sí. Ambos supuestos son falsos, sin embargo permiten simplificar mucho los cálculos y obtener de igual forma resultados muy buenos.

Las probabilidades se calculan con los estimadores de máxima verosimilitud -las frecuencias relativas- de los datos de entrenamiento. Una vez estimados los parámetros, se calculan las probabilidades a posteriori de pertenencia a las distintas clases, y cada caso es finalmente asignado a la clase más probable.

Los clasificadores Bayesianos se aplican mejor a problemas en los que la información de numerosos atributos se debe considerar simultáneamente para estimar la probabilidad general de un resultado. Si bien muchos algoritmos de aprendizaje automático ignoran las características que tienen efectos débiles, los métodos bayesianos utilizan toda la evidencia disponible para cambiar sutilmente las predicciones. Si gran cantidad de características tienen efectos relativamente menores, en conjunto, su impacto combinado podría ser bastante grande.

Dos de las ventajas que presenta el método son su simplicidad y eficiencia computacional. A pesar de sus inconvenientes, este método ha sido muy utilizado históricamente y se han obtenido buenos resultados

para muchos conjuntos de datos. Esto ocurre cuando el conjunto de entrenamiento representa bien las distribuciones de probabilidad del problema.

Aplicación

Utilizaremos para la aplicación de este método la función `naiveBayes()` del paquete `e1071` de R. Debe indicarse el conjunto de predictores x y la variable de clasificación y, o de forma alternativa una fórmula como se verifica a continuación.

```
set.seed(123)
modelo_naive <- naiveBayes(ACTIVOS ~ ., data = activos_train)
```

El objeto construido "modelo_naive" devuelve la distribución de probabilidades a priori de las dos clases y para cada predictor las probabilidades condicionadas por cada clase.

```
modelo_naive
```

```
  Naive Bayes Classifier for Discrete Predictors

  Call:
  naiveBayes.default(x = X, y = Y, laplace = laplace)

  A-priori probabilities:
  Y
  ocupados  parados
    0.8634   0.1366

  Conditional probabilities:
            SEXO
  Y          hombre  mujer
    ocupados 0.5317 0.4683
    parados  0.4517 0.5483

            EDAD
  Y           16 a 19  20 a 24  25 a 29  30 a 34  35 a 39  40 a 44  45 a 49
    ocupados 0.009341 0.045854 0.069138 0.086022 0.118821 0.153709 0.154657
    parados  0.044121 0.124053 0.108164 0.097898 0.094109 0.122586 0.119653
            EDAD
  Y           50 a 54  55 a 59  60 a 64 65 o más
    ocupados 0.150866 0.125841 0.072813 0.012938
    parados  0.122586 0.106087 0.057688 0.003055

            ECIVIL
  Y           Casado Sep/divorciado Soltero   Viudo
    ocupados 0.57185        0.07771 0.33701 0.01344
    parados  0.39440        0.08898 0.50452 0.01210

            ESTU
  Y                         AN       P1       P2       S1       SG       SP
    ocupados 0.000000 0.001412 0.006923 0.040535 0.279899 0.130541 0.104955
    parados  0.000000 0.006355 0.021633 0.093865 0.378147 0.133464 0.108531
            ESTU
  Y                SU
```

```
    ocupados 0.435735
    parados  0.258005

             NAC
  Y           Doble nac. Española Extranjera
    ocupados    0.02883  0.90367    0.06749
    parados     0.03850  0.84466    0.11684

             CCAA
  Y           Andalucía   Aragón Asturias Canarias Cantabria Castilla-La Mancha
    ocupados  0.142222 0.047072 0.026166 0.037770  0.023807           0.067301
    parados   0.253483 0.029455 0.028477 0.064410  0.014666           0.076265
             CCAA
  Y           Castilla y León Cataluña    Ceuta Comunitat Valenciana Extremadura
    ocupados        0.094995 0.110989 0.003500             0.076874    0.029783
    parados         0.070032 0.080665 0.009289             0.074065    0.047543
             CCAA
  Y            Galicia Illes Balears La Rioja   Madrid  Melilla   Murcia  Navarra
    ocupados 0.117081      0.030305 0.018817 0.063743 0.003288 0.032742 0.024484
    parados  0.093376      0.016133 0.011611 0.046199 0.008066 0.033733 0.012711
             CCAA
  Y           País Vasco
    ocupados   0.049064
    parados    0.029822
```

A continuación se lleva a cabo la predicción del conjunto de prueba y se calcula la matriz de confusión, cruzando la variable de predicción con la clase original. En la función `confusionMatrix()` se establece como clase positiva el grupo de parados.

Establecemos la predicción mediante la función `predict()`, creando el objeto "pred_naive".

```
pred_naive <- predict(modelo_naive, x_test, type = "class")
```

```
confusion_naive <- confusionMatrix(pred_naive, y_test, positive = "parados")
confusion_naive
```

```
  Confusion Matrix and Statistics

            Reference
  Prediction ocupados parados
    ocupados    12713    1877
    parados       214     168

                 Accuracy : 0.86
                   95% CI : (0.855, 0.866)
      No Information Rate : 0.863
      P-Value [Acc > NIR] : 0.866

                    Kappa : 0.1

   Mcnemar's Test P-Value : <2e-16

              Sensitivity : 0.0822
```

```
            Specificity : 0.9834
         Pos Pred Value : 0.4398
         Neg Pred Value : 0.8714
             Prevalence : 0.1366
         Detection Rate : 0.0112
   Detection Prevalence : 0.0255
      Balanced Accuracy : 0.5328

       'Positive' Class : parados
```

Las variables explicativas permiten predecir a los parados en tan solo el 8.22 % de los casos, se obtiene un bajo porcentaje de aciertos. Si queremos saber la eficiencia del clasificador al separar ambas clases miramos el *accuracy* balanceado (53.28 %), que es más recomendable en estos casos donde las clases están desbalanceadas.

6.3.2. CART

CART (Árboles de clasificación y regresión, Breiman et al.1984) es un método que construye árboles binarios con los casos de la muestra.

El árbol de decisión es una técnica clásica de clasificación. La idea viene a través de la conocida estructura de los árboles, donde podemos encontrar nodos, ramas y hojas.

Se denomina también "segmentación jerárquica", ya que utiliza un proceso de división secuencial, iterativo y descendente donde se emplean particiones binarias de la variable dependiente donde en cada iteración se selecciona una variable independiente y un punto de corte que minimiza un error de clasificación. Esta variable independiente va a representar la variable de mayor relevancia, siendo aquel atributo que tenga la mayor ganancia de información en el proceso de clasificación.

El proceso que realiza es el siguiente:

- Se comienza con un nodo inicial. Se elige la variable x_1 y se determina un punto de corte, por ejemplo, c de modo que se puedan separar los datos en dos conjuntos que serán las ramas: aquellos con $x_1 \leq c$ y los que tienen $x_1 > c$, siendo las ramas aquellas que representan los posibles valores de las variables de entrada.
- De este nodo inicial saldrán ahora dos: uno al que llegan las observaciones con $x_1 \leq c$, y otro al que llegan las observaciones con $x_1 > c$.

Hacemos esto de manera recursiva hasta que el algoritmo determina que los datos dentro de los subconjuntos son suficientemente homogéneos, y que no es posible obtener una mejor separación. Cuando esto ocurre el algoritmo se detiene.

Se le llama nodo terminal u hoja a los posibles valores de la variable de salida, corresponden a una clase o una etiqueta.

Cuando CART se utiliza como método de clasificación, la homogeneidad de los grupos formados en la división del árbol se traduce en que dentro de cada grupo predominen los elementos pertenecientes a una sola clase, lo que se conoce como "pureza" del nodo. Se han propuesto dos medidas del grado de pureza: el índice de Gini, y la entropía cruzada, también llamada devianza o cantidad de información. En cada paso de la construcción del árbol se mejora la pureza global minimizando el valor del estadístico de ajuste.

Para dos clases, el índice de Gini para cada grupo formado se define como:

$$G = p_1(1 - p_1) + p_2(1 - p_2)$$

donde p_1 y p_2 son las probabilidades de ambas clases. Como ambos valores suman la unidad $p_1 = 1 - p_2$, y el índice se puede expresar como $G = 2p_1p_2$ cuyo valor es cero cuando alguna de las probabilidades es cero, es decir cuando el grupo es puro (homogéneo) y solo contiene elementos de una clase. El valor máximo se obtiene cuando $p_1 = p_2 = 0,5$ indicando el mayor grado de impureza. El índice de Gini para dos grupos se calcula como la media de ambos ponderada con el número de casos.

La segunda medida de pureza, la entropía cruzada se define como

$$I = -[p_1 \log_2(p_1) + p_2 \log_2(p_2)]$$

cuyo valor se sustituye por cero cuando alguna de las probabilidades es nula. El valor de I, al igual que el índice de Gini, es más pequeño cuanto mayor es la pureza de los grupos.

Los árboles de decisión son uno de los métodos de aprendizaje inductivo supervisado no paramétrico más utilizado. Son apropiados cuando hay un número elevado de datos. Una de sus ventajas su carácter descriptivo que permite entender e interpretar fácilmente las decisiones tomadas por el modelo. Este es un algoritmo que no es demandante en poder de cómputo, comparado con procedimientos más sofisticados, y a pesar de ello tiende a arrojar buenos resultados de predicción para muchos tipos de datos.

Como debilidades en este tipo de algoritmos está el bajo porcentaje de precisión si lo comparamos con otros métodos de aprendizaje automático. Su tipo de clasificación es débil, ya que sus resultados pueden variar mucho dependiendo de la muestra de datos utilizados para entrenar un modelo.

El árbol de clasificación completo tiene un claro problema de sobreajuste: en general clasifica muy bien los elementos de la muestra, pero la precisión puede disminuir drásticamente cuando se predicen casos nuevos, datos no utilizados para la construcción del árbol. Para reducir ese problema se utiliza un procedimiento de pruning o "poda" del árbol que penaliza la complejidad y lo reduce, eliminando elementos superfluos o criterios irrelevantes. Cuanto mayor es el tamaño del árbol mayor es también la precisión o pureza conseguida (con los datos de la muestra), pero también es mayor el sobreajuste. Se trata de reducir el tamaño manteniendo una precisión aceptable, buscando un equilibrio razonable entre tamaño y precisión. En general un árbol pequeño es más sencillo, más fácil de interpretar, y tiene menos problema de sobreajuste.

Dicho esto, en caso de que el árbol sea grande se suele "podar" para disminuir su complejidad, y aumentar la precisión de la predicción. Existen dos formas de poda utilizadas: la poda por coste complejidad y la poda pesimista. En la poda por coste complejidad se trata de equilibrar la precisión y el tamaño del árbol. La complejidad está determinada por el número de hojas que posee el árbol. La poda pesimista utiliza los casos clasificados incorrectamente y obtiene un error de sustitución, eliminando los sub-árboles que no mejoran significativamente la precisión del clasificador.

En la opción "control" de la función `rpart()`, encontramos parámetros opcionales para controlar el crecimiento del árbol. El principal parámetro de esta función es la complejidad, `cp`. Otros parámetros importantes son `minsplit`, número mínimo de observaciones que debe haber en un nodo para que se intente una partición, y 'minbucket', número mínimo de observaciones de un nodo terminal. Por defecto `minsplit=20, minbucket=round(minsplit/3)` y `cp=0,01`.

Aplicación

Determinar parámetros

Antes de emplear el modelo, decidimos determinar los parametros a aplicar. El procedimiento habitual para determinar los parámetros del modelo consiste en dejar que el árbol crezca hasta que cada nodo terminal tenga menos de un número mínimo de observaciones, ya que en algunas divisiones puede que apenas mejore y en las siguientes sí lo haga. Para determinar "cp" se considera como 0, para que el árbol crezca lo máximo posible.

Como el árbol que se obtiene es muy grande, se podaría considerando una función de coste basada en la complejidad. Para esto, se calcula el error de la validación cruzada (Figura 6.1) y se elige el menor.

```
set.seed(123)

rpart <- rpart(ACTIVOS ~ ., data = activos_train, method = "class",
    parms = list(split = "information"), control = rpart.control(minsplit = 20,
        minbucket = 10, cp = 0, usesurrogate = 0, maxsurrogate = 0))
```

```
plotcp(rpart)
```

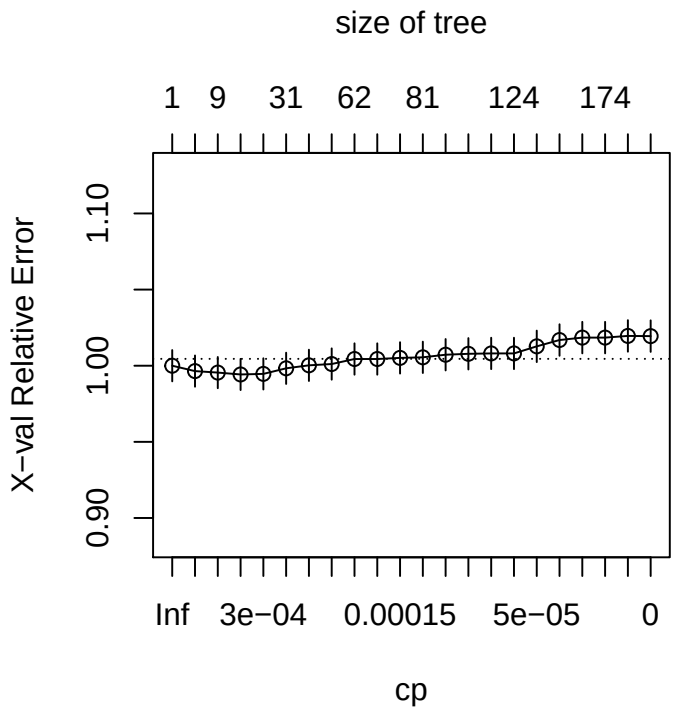


Figura 6.1: Gráfico del error de la validación cruzada.

```
printcp(rpart)
```

```
Classification tree:
rpart(formula = ACTIVOS ~ ., data = activos_train, method = "class",
    parms = list(split = "information"), control = rpart.control(minsplit = 20,
        minbucket = 10, cp = 0, usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] CCAA   ECIVIL EDAD   ESTU   NAC    SEXO

Root node error: 8182/59890 = 0.14

n= 59890

        CP nsplit rel error xerror xstd
1  1.1e-03      0      1.00   1.00 0.01
```

```
   2  8.6e-04      7      0.99   1.00 0.01
   3  4.9e-04      8      0.99   1.00 0.01
   4  3.1e-04     13      0.99   0.99 0.01
   5  2.9e-04     24      0.98   0.99 0.01
   6  2.4e-04     30      0.98   1.00 0.01
   7  2.2e-04     46      0.98   1.00 0.01
   8  2.0e-04     53      0.97   1.00 0.01
   9  1.8e-04     61      0.97   1.00 0.01
   10 1.6e-04     65      0.97   1.00 0.01
   11 1.5e-04     75      0.97   1.01 0.01
   12 1.2e-04     80      0.97   1.01 0.01
   13 9.2e-05     96      0.97   1.01 0.01
   14 8.7e-05    104      0.97   1.01 0.01
   15 8.1e-05    114      0.97   1.01 0.01
   16 6.1e-05    123      0.97   1.01 0.01
   17 4.1e-05    144      0.96   1.01 0.01
   18 3.5e-05    150      0.96   1.02 0.01
   19 3.1e-05    157      0.96   1.02 0.01
   20 2.4e-05    173      0.96   1.02 0.01
   21 1.5e-05    188      0.96   1.02 0.01
   22 0.0e+00    204      0.96   1.02 0.01
```

Tal y como se puede ver en la tabla anterior, el mínimo error se alcanza en el nodo 4. Típicamente se considera que hasta la línea discontinua de la Figura 6.1 no hay diferencias significativas. Esta línea es la suma del mínimo error y la desviación típica, es decir:

```
0.99 + 0.01
```

```
[1] 1
```

Se toma un modelo más simple, que en este caso es el 1, de cp = 0.0011 y se construye el árbol final que se puede ver en la Figura 6.2.

```
set.seed(123)
modelo_cart <- rpart(ACTIVOS ~ ., data = activos_train, method = "class",
    control = rpart.control(minsplit = 20, minbucket = 10, cp = 0.0011))
```

Del entrenamiento de nuestro modelo obtenemos el esquema de nuestro árbol de clasificación. Cada inciso nos indica un nodo y la regla de clasificación que le corresponde. Especificamente, muestra para cada nodo la variable utilizada y el punto de corte, el número de casos de cada clase, y la proporción o probabilidad correspondiente.

Siguiendo estos nodos podemos llegar a las hojas del árbol, correspondientes con la clasificación de nuestros datos. Algunos nodos como el 2, 6, 17, están marcados con un asterisco (*), lo que indica que son nodos terminales, no se dividen más.

```
print(modelo_cart)
```

```
n= 59890

node), split, n, loss, yval, (yprob) * denotes terminal node

  1) root 59890 8182 ocupados (0.8634 0.1366)
```

2) EDAD=25 a 29,30 a 34,35 a 39,40 a 44,45 a 49,50 a 54,55 a 59,60 a 64,65 o más 55660 6806 ocupados (0.8777 0.1223) *
3) EDAD=16 a 19,20 a 24 4230 1376 ocupados (0.6747 0.3253)
   6) CCAA=Aragón,Cantabria,Castilla y León,Cataluña,Comunitat Valenciana,Galicia,Illes Balears,La Rioja,Madrid,Murcia,Navarra,País Vasco 2794 739 ocupados (0.7355 0.2645) * 7) CCAA=Andalucía,Asturias,Canarias,Castilla-La Mancha,Ceuta,Extremadura,Melilla 1436 637 ocupados (0.5564 0.4436)
      14) EDAD=20 a 24 1165 483 ocupados (0.5854 0.4146)

      15) SEXO=hombre 652 237 ocupados (0.6365 0.3635)
         56) CCAA=Andalucía,Asturias,Canarias,Castilla-La Mancha,Extremadura 633 222 ocupados (0.6493 0.3507) *57) CCAA=Ceuta,Melilla 19 4 parados (0.2105 0.7895)*
      16) SEXO=mujer 513 246 ocupados (0.5205 0.4795)
         58) ESTU=SG,SP,SU 374 161 ocupados (0.5695 0.4305) *59) ESTU=P1,P2,S1 139 54 parados (0.3885 0.6115)*
      17) EDAD=16 a 19 271 117 parados (0.4317 0.5683) *

Todo lo anterior resulta mucho más claro si lo visualizamos en una gráfica representando el modelo con la función `rpart.plot()`, del paquete del mismo nombre.

```
rpart.plot(modelo_cart)
```

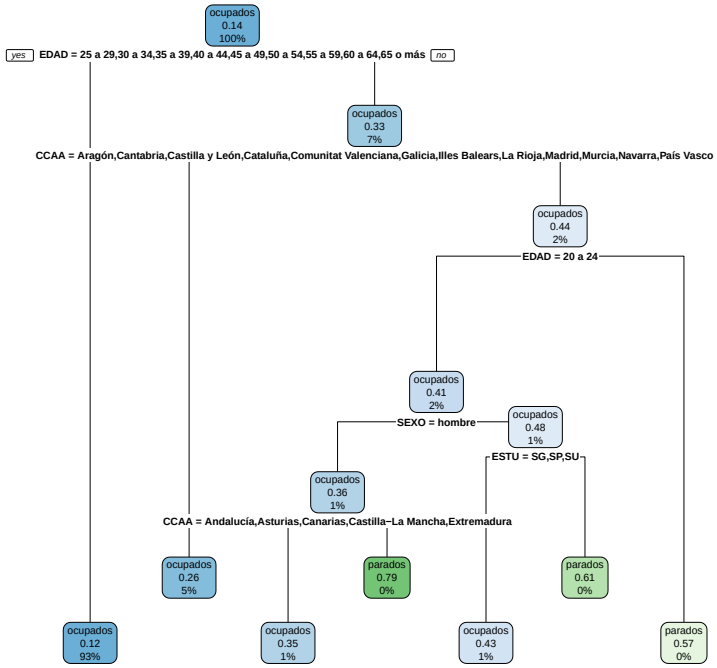


Figura 6.2: Árbol de decisión.

En estos gráficos cada uno de los cubos representa un nodo de nuestro árbol con su regla de clasificación. La idea es que para la clasificación de un nuevo ejemplo se estudia a qué nodo terminal del árbol pertenece.

En la Figura 6.2, el nodo 1 (raiz) indica que de 59,890 casos, el 14 % se encuentran parados y el restante 86 % ocupados. Los nodos terminales –hojas del árbol- muestran la clase predominante (Ocupado/Parado),

la proporción de casos de Activos = Parados dentro del nodo, y el porcentaje de casos del nodo sobre el total de la muestra. Los nodos terminales se podría tomar como una estimación de las probabilidades de que pertenezca a cada categoría.

El nodo raiz prueba el atributo `EDAD` si la persona tiene de 25 años o más (conformando estos el 93 % de la muestra), nos desplazamos por las ramas izquierdas del árbol, el arbol clasifica a estos individuos como ocupados, con una probabilidad estimada del 88 %.

En caso de que tengan menos de 25 años, nos desplazamos por las ramas de la derecha del árbol. Aquí vuelve y hace una nueva división utilizando la variable `CCAA`.

Si los individuos pertenecen a alguna de las comunidades autónomas especificadas nos desplazamos por la rama izquierda estimando como parados al 26 %. De los 1436 restantes, nuevamente el clasificador utiliza la variable `EDAD` para asignarle la actividad económica a la que pertenece. Si estos jóvenes tienen menos de 20 años pues estima que el 57 % de estos esta parados. Mientra que si tienen de 20 a 24 años nos desplazamos hacia la izquierda y continua la división, en este próximo caso utilizando la variable `SEXO`. Así continúa hasta llgar a los nodos terminales.

A continuación, se lleva a cabo la predicción del conjunto de prueba y se calcula la matriz de confusión.

Para cada uno de los datos del conjunto de prueba se sigue el árbol según el valor de sus variables hasta llegar a los nodos terminales, allí, se clasifica con la categoría más probable de ese nodo terminal, como se observó en el árbol de la Figura 6.2.

```
pred_cart <- predict(modelo_cart, x_test, type = "class")
```

```
confusion_cart <- confusionMatrix(pred_cart, y_test, positive = "parados")
```

```
confusion_cart
```

```
  Confusion Matrix and Statistics

            Reference
  Prediction ocupados parados
    ocupados    12873    1984
    parados        54      61

                 Accuracy : 0.864
                   95% CI : (0.858, 0.869)
      No Information Rate : 0.863
      P-Value [Acc > NIR] : 0.44

                    Kappa : 0.043

   Mcnemar's Test P-Value : <2e-16

              Sensitivity : 0.02983
              Specificity : 0.99582
           Pos Pred Value : 0.53043
           Neg Pred Value : 0.86646
               Prevalence : 0.13659
           Detection Rate : 0.00407
     Detection Prevalence : 0.00768
        Balanced Accuracy : 0.51283
```

```
'Positive' Class : parados
```

Se obtiene un AUC de 51.28 %, menor que la obtenida por el modelo de *naive*.

El poder de predicción de los árboles de decisión no suele ser muy bueno, pero el algoritmo es sencillo y los modelos resultantes tienen una fácil interpretación. Con el fin de mejorar esta predicción se puede seguir la idea del *bagging* de combinar muchos métodos sencillos, como se verá en el método de bosques aleatorios.

6.3.3. Algoritmo K-vecinos más cercanos

El clasificador k vecinos más próximos, *k-nn* (Cover y Hart, 1967) estima el valor de la función de densidad $F(X/C_i)$ de probabilidad de que un elemento X pertenezca a la clase C_i a partir de la información proporcionada por la muestra de aprendizaje. Es un método de clasificación supervisada basado en una idea muy simple e intuitiva: cada elemento es asignado a la misma clase que sus vecinos más cercanos. Se trata de un método no paramétrico, en el que no se hace ningún supuesto a priori sobre la distribución.

Los casos de entrenamiento son vectores en un espacio de dimensión m, descrito en términos de m atributos o variables explicativas, suponiendo conocida la clase a la que pertenece cada uno de los casos. Para la asignación de cada caso (punto o vector en ese espacio m-dimensional) se consideran los k elementos de la muestra más próximos a ese punto -los k vecinos más cercanos- y se asigna el caso a la clase C más frecuente entre ellos. Generalmente se utiliza la distancia euclídea, esta distancia se mide en línea recta utilizando la ruta directa más corta.

El valor k será el número de vecinos más cercanos con características similares para realizar la comparación y llevar a cabo la elección.

El algoritmo es el siguiente:

1) Para cada caso i que debe ser asignado se calcula su distancia a cada uno de los casos ya asignados.

2) Se seleccionan los k casos más próximos a i.

3) Se asigna i a la clase más frecuente dentro del conjunto de k casos más próximos.

Si el número de atributos es grande puede ocurrir que muchos de ellos sean irrelevantes, pero predominan en la asignación, ya que son mayoría, sobre los atributos importantes. Para corregir este efecto se identifican y eliminan previamente las variables superfluas, o bien se asigna un peso diferente a cada una en el cálculo de las distancias, reajustando los pesos durante el entrenamiento. Para ello se puede utilizar como ponderación la información mutua media $I(Xi, C)$ entre la variable i y la clase C, que mide la información que comparten X_i y C, e indica hasta que punto el conocimiento de una de ellas reduce nuestra incertidumbre sobre la otra. Si X_i y C son independientes, en cuyo caso X_i es irrelevante, la información mutua será cero; cuanto mayor sea la relación entre ambas, mayor será la información mutua.

También es posible atribuir un mayor peso en el cálculo de las frecuencias a los casos más próximos, utilizando por ejemplo el inverso del cuadrado de la distancia como factor de ponderación.

La mejor elección de k depende de los datos y su aplicación. La decisión de cuántos vecinos usar para *k-nn* determina qué tan bien se generalizará el modelo para datos futuros. El equilibrio entre el ajuste excesivo y el ajuste insuficiente de los datos de entrenamiento es un problema conocido como compensación de sesgo-varianza. Generalmente valores grandes de k reducen el efecto del ruido en la clasificación, pero puede sesgar el algoritmo y separar clases relativamente próximas.

A pesar de la simplicidad de esta idea, los métodos del vecino más cercano son muy potentes. La principal ventaja del algoritmo *k-nn* es su facilidad de implementación, aunque su costo computacional es alto a medida que aumenta el tamaño de los datos usados en el entrenamiento.

Aplicación

Con la función `knn()` del paquete `class` se entrena el método de los k vecinos más próximos. La función emplea como criterio la distancia euclídea, y decide la asignación por mayoría entre los k vecinos más próximos (al azar en caso de empate). Si hay varios vecinos a la misma distancia que el k-ésimo se incluyen todos ellos.

La función toma el conjunto de entrenamiento y de *test*, así como la variable que contiene la clase de `Activos`, devolviendo la predicción de clasificación del *test*. Por defecto toma $k = 1$.

```
set.seed(123)

x <- model.matrix(ACTIVOS ~ ., data = activos_train)[, -1]
x2 <- model.matrix(ACTIVOS ~ ., data = activos_test)[, -1]
modelo_knn <- knn(x, x2, y_train)
```

Procedemos a realizar la evaluación del modelo.

```
confusion_knn <- confusionMatrix(modelo_knn, y_test, positive = "parados")
confusion_knn
```

```
  Confusion Matrix and Statistics

            Reference
  Prediction ocupados parados
    ocupados    12516    1868
    parados       411     177

                 Accuracy : 0.848
                   95% CI : (0.842, 0.854)
      No Information Rate : 0.863
      P-Value [Acc > NIR] : 1

                    Kappa : 0.078

   Mcnemar's Test P-Value : <2e-16

              Sensitivity : 0.0866
              Specificity : 0.9682
           Pos Pred Value : 0.3010
           Neg Pred Value : 0.8701
               Prevalence : 0.1366
           Detection Rate : 0.0118
     Detection Prevalence : 0.0393
        Balanced Accuracy : 0.5274

         'Positive' Class : parados
```

El *accuracy* balanceado es de 52.74 %. Con ún 8.66 % de aciertos de la clase de parados.

6.3.4. Random Forest

Random Forest o Bosque Aleatorio (Leo Breiman y Adele Cutler, 2001) que combina las técnicas CART y *Bagging*. Tiene como propósito incorporar la aleatoriedad en las distintas etapas de la construcción de

un árbol obtenido por CART. Se trata de una técnica de clasificación supervisada no paramétrica que busca segmentar el espacio de los predictores en regiones simples, definidas por intervalos de valores de los predictores o variables explicativas, dentro de las cuales la predicción es mucho más eficaz que en el espacio completo.

En este tipo de algoritmos se realiza una búsqueda de un clasificador más preciso con el uso de árboles de decisión. El bosque aleatorio crea múltiples árboles de decisión, utilizando el concepto denominado *bagging*, para introducir un muestreo aleatorio en todo el proceso. Después de generar el conjunto de árboles (el bosque), el modelo combina las predicciones de los árboles.

La técnica *Bagging* fue introducida por Breiman en 1996, es un método de agregación de modelos homogéneos que se basa sobre el voto mayoritario o el promedio según el caso. Es una técnica usada para reducir la varianza de las predicciones a través de la combinación de los resultados de varios clasificadores. Cada uno de ellos modelados con diferentes subconjuntos tomados de la misma población. La selección de estas muestras denominadas *bag* se toma aleatoriamente con reemplazamiento. Si existen M variables de entrada, un número $m < M$ se especifica tal que para cada nodo, m variables se seleccionan aleatoriamente de M. La mejor división de estos m atributos es usado para ramificar el árbol. El valor m se mantiene constante durante la generación de todo el bosque.

Random Forest ajusta un número elevado de árboles, en lugar de uno solo. Cada árbol da una clasificación y tiene un peso equivalente en el proceso de toma de decisiones final. La clasificación de un nuevo ejemplo será la predicción más frecuente de estos árboles.

Al considerar solo una pequeña fracción del número total de variables disponibles, la cantidad de memoria requerida se reduce significativamente. Cada árbol crece hasta su máxima extensión posible y no hay proceso de poda, como ocurre en un árbol de decisión simple.

La diferencia con un árbol de decisión es que en el nodo donde tenemos que hacer la división de los datos no se toma aquel atributo con mayor ganancia o índice *Gini*, sino que se coge la mejor división entre un subconjunto aleatorio de todo el conjunto de características. Esto provocará un sesgo mayor del árbol, sin embargo al usar la media de varios árboles la varianza disminuye.

En cada división del árbol elige las variables más adecuadas -buscando el mejor resultado- y su punto de corte, utilizando para ello un número de variables candidatas igual a la raíz cuadrada de m, el número de predictores.

La función que utilizaremos con R para este método es `randomForest()`, del paquete del mismo nombre.

Aplicación

Para este algoritmo utilizaremos una muestra de 10,000 observaciones, por el alto tiempo de cómputo.

```
set.seed(123)
modelo_rf <- randomForest(ACTIVOS ~ ., data = activos_train)
```

```
modelo_rf
```

```
  Call:
   randomForest(formula = ACTIVOS ~ ., data = activos_train)
                 Type of random forest: classification
                       Number of trees: 500
  No. of variables tried at each split: 2

          OOB estimate of  error rate: 14.14%
  Confusion matrix:
           ocupados parados class.error
```

```
ocupados     6824      75     0.01087
parados      1056      45     0.95913
```

El número de árboles que construye el bosque es 500 por defecto, número considerado adecuado para la mayoría de las aplicaciones, aunque puede ser cambiado con el argumento opcional `ntree`. El modelo muestreó 2 predictores aleatorios en cada división.

El error *out-of-bag* (OOB), que mide el error de predicción, es de 14.14 %. También muestra una matriz que contiene los valores predichos versus los reales, así como el error de clasificación para cada clase en la muestra de aprendizaje.

Variables importantes

El método calcula una medida de la importancia de cada variable explicativa en el proceso de clasificación, basada en el índice de Gini, que mide el grado de pureza o de homogeneidad de los nodos del árbol: un nodo es puro cuando solo tiene elementos de la misma clase; el nodo es más puro cuanto más pequeño es el valor del índice. La medida de importancia consiste en la reducción del índice de *Gini* debida a esa variable, por lo que las variables más importantes son las que consiguen reducción mayor.

Se puede ver con el elemento de resultado `rf_model$importance`:

```
modelo_rf$importance
```

```
       MeanDecreaseGini
SEXO              42.00
EDAD             179.05
ECIVIL            66.32
ESTU             114.83
NAC               48.04
CCAA             220.43
```

Las variables más importantes en la identificación de parados/ocupados son `CCAA` (220.43) y `EDAD` (179.05). De acuerdo con estos resultados `SEXO` es la variable de menos importancia.

Veamos ahora la identificación de los `Activos` para los casos que hemos mantenido al margen del modelo, con el fin de ser utilizados como validación, en el conjunto de prueba:

```
pred_rf <- predict(modelo_rf, newdata = x_test)
```

A continuación cruzamos la variable de predicción con la clase original para conocer el grado de acierto.

```
confusion_rf <- confusionMatrix(pred_rf, y_test2, positive = "parados")
```

```
confusion_rf
```

```
Confusion Matrix and Statistics

          Reference
Prediction ocupados parados
  ocupados     1713     264
  parados        12      11

               Accuracy : 0.862
                 95% CI : (0.846, 0.877)
```

```
      No Information Rate : 0.863
      P-Value [Acc > NIR] : 0.542

                    Kappa : 0.054

   Mcnemar's Test P-Value : <2e-16

              Sensitivity : 0.0400
              Specificity : 0.9930
           Pos Pred Value : 0.4783
           Neg Pred Value : 0.8665
               Prevalence : 0.1375
           Detection Rate : 0.0055
     Detection Prevalence : 0.0115
        Balanced Accuracy : 0.5165

         'Positive' Class : parados
```

A pesar de ser un modelo con el que se suelen obtener buenos resultados, tan solo acierta el 4 % de los parados con un accuracy balance de 51.65 %.

6.3.5. Adaboost

AdaBoost (*Adaptive Boosting*, refuerzo adaptativo, Freund y Schapire, 1996) es un meta-algoritmo de aprendizaje automático, y se puede utilizar para aumentar el rendimiento de otros algoritmos de aprendizaje. La técnica de modelización *Boosting* es una de las más potentes introducidas en los últimos 20 años. Se diseñó en el año 1989 de la mano del profesor Robert Schapire, aunque un año después se mejora por Jonh E. Freund. Originalmente se creó para abordar problemas de clasificación, pero se llegó a extender también para la regresión.

El "refuerzo adaptativo" consiste en una estrategia que pondera de forma distinta los casos bien y mal clasificados en cada iteración, dando un peso mayor a los mal clasificados, para modificar de forma progresiva la regla de decisión.

Adaboost construye de forma iterativa un conjunto de reglas simples de clasificación (clasificadores débiles), que después se utilizan de forma conjunta para elaborar una regla compleja (clasificador fuerte), de forma similar a la utilización de un conjunto de expertos que decide por mayoría -para resolver un problema determinado- en lugar de emplear un único experto.

Un clasificador aleatorio decide al azar la clase a la que pertenece cada elemento. Se denomina clasificador débil a aquel que produce resultados solo ligeramente mejores que un clasificador aleatorio. En cualquier conjunto de datos tendremos en general muchas variables relacionadas con el criterio de clasificación o variable de interés, y la mayoría de ellas pueden ser utilizadas como clasificador débil.

A diferencia del *Bagging*, no se crean versiones del conjunto de entrenamiento, sino que se trabaja secuencialmente siempre con el mismo conjunto de entrada manipulando los pesos de los datos para generar modelos distintos. Es decir, consiste en ir ajustando de manera individual B árboles, los cuáles presentan mucho sesgo pero poca varianza, de modo que cada árbol nuevo no cometa los errores del árbol anterior, y así mejorar iteración tras iteración.

Si buscamos una clasificación con k clases, y decidimos la clase de cada elemento de forma totalmente aleatoria acertaremos en promedio $1/k$ (50 % si son dos clases), y el error de clasificación será $1 - 1/k$: a los clasificadores débiles se les pide simplemente que tengan un error menor que $1 - 1/k$, y del resto se encarga *AdaBoost*.

Normalmente los Árboles de decisión se utilizan como clasificadores débiles construyendo una estructura con pocas ramificaciones, por ello, recuperando la notación utilizada hasta ahora del número de árboles (B), el algoritmo *AdaBoost* se estructuraría de la siguiente manera:

1) Se asignan pesos o ponderaciones iniciales iguales $1/N$ a todos los casos de la muestra de entrenamiento. Donde N es el tamaño de la muestra.

2) Se aplica el clasificador débil a los datos ponderados, y se asigna al clasificador un peso en función de la tasa de error "e" obtenida: $peso = 05ln[(1-e)/e]$. Este paso se repite T veces modificando los pesos de los casos, aumentando los correspondientes a los casos mal asignados, con el fin de centrar los esfuerzos en clasificar bien los casos erróneos. En cada iteración se recalculan las ponderaciones de los casos, se aplica el clasificador, y se calcula el peso o ponderación del clasificador obtenido.

3) Finalmente se construye el clasificador mejorado como una combinación lineal de los T clasificadores encontrados, aplicando los pesos calculados en las T iteraciones.

El *AdaBoost* es sensible al ruido en los datos y casos aislados, sin embargo, se considera que es menos susceptible al sobreajuste, problema común a la mayoría de los métodos de clasificación supervisada.

Entre las ventajas se puede encontrar que es simple y sencillo de programar, el único parámetro a establecer son las iteraciones. El clasificador débil no requiere conocimiento previo, es versátil y rápido.

Por otro lado, *boosting* a veces conduce a un deterioro en el rendimiento de generalización. Según Quinlan (1996), la razón principal para el fracaso de *boosting* es el sobreajuste. El objetivo de este algoritmo es construir un clasificador compuesto que funcione bien en los datos, pero un gran número de iteraciones puede crear un clasificador compuesto muy complejo, que es significativamente menos preciso que un solo clasificador. Una posible forma de evitar el sobreajuste es mantener el número de iteraciones lo más pequeño posible.

Otro inconveniente importante es que es difícil de entender. El conjunto resultante se considera menos comprensible ya que el usuario debe capturar varios clasificadores en lugar de un solo clasificador. A pesar de los inconvenientes anteriores, Breiman (1996) se refiere a la idea de *boosting* como el desarrollo más significativo en el diseño de clasificadores de los años noventa.

Aplicación

Utilizaremos para la aplicación de este método la función `boosting()`, del paquete de R `adabag`. Esta función realiza el refuerzo adaptativo empleando como clasificadores débiles árboles de clasificación, que son entrenados con los datos de la muestra.

Para este algoritmo utilizaremos una muestra de 10,000 observaciones, por el alto tiempo de cómputo.

```
set.seed(123)
modelo_adaboost <- boosting(ACTIVOS ~ ., data = activos_train)
```

El algoritmo genera 100 reglas o clasificadores débiles con forma de árbol de decisión, que se pueden listar con:

```
modelo_adaboost$trees
```

El último árbol, número 100, se muestra a continuación a título de ejemplo.

```
modelo_adaboost$trees[[100]]
```

n= 8000

node), split, n, loss, yval, (yprob) * denotes terminal node

1) root 8000 3982 ocupados (0.5022 0.4978)

2) CCAA=Andalucía,Aragón,Asturias,Canarias,Cantabria,Castilla-La Mancha,Castilla y León,Extremadura,Galicia,Illes Balears,Murcia,Navarra,País Vasco 5992 2917 ocupados (0.5132 0.4868) *
3) CCAA=Cataluña,Ceuta,Comunitat Valenciana,La Rioja,Madrid,Melilla 2008 943 parados (0.4696 0.5304) *

La asignación final se realiza con las 100 reglas o clasificadores conjuntamente, ponderando cada una de ellas con el peso asignado al clasificador.

Variables importantes

A continuación mostramos la importancia de cada variable en el proceso de asignación.

```
modelo_adaboost$importance
```

```
   CCAA ECIVIL   EDAD   ESTU    NAC   SEXO
37.763 14.283 19.165 20.039  4.719  4.031
```

El estadístico de importancia está expresado en porcentaje (la suma es 100).Las variables `CCAA` (37.76 %), `ESTU` (20,03 %) y `EDAD` (19.16 %) son las que contribuyen en mayor medida a la clasificación.

Confirmamos si el método servirá para clasificar elementos nuevos, no utilizados en el análisis. Para esto lo aplicamos a la muestra de prueba.

```
pred_adaboost <- predict.boosting(object = modelo_adaboost, newdata = x_test,
    type = "class")
```

```
confusion_adaboost <- confusionMatrix(as.factor(pred_adaboost$class),
    y_test2, positive = "parados")
confusion_adaboost
```

```
  Confusion Matrix and Statistics

            Reference
  Prediction ocupados parados
    ocupados     1708     258
    parados        17      17

                 Accuracy : 0.863
                   95% CI : (0.847, 0.877)
      No Information Rate : 0.863
      P-Value [Acc > NIR] : 0.516

                    Kappa : 0.082

   Mcnemar's Test P-Value : <2e-16

              Sensitivity : 0.0618
              Specificity : 0.9901
           Pos Pred Value : 0.5000
           Neg Pred Value : 0.8688
               Prevalence : 0.1375
```

```
      Detection Rate : 0.0085
Detection Prevalence : 0.0170
   Balanced Accuracy : 0.5260

    'Positive' Class : parados
```

Al cruzar la clase real con la predicción observamos que el porcentaje de acierto es de 86.3 %; el método identifica con precisión la efectividad de empleo.

Cuando evaluamos los indicadores *accuracy balance*, *Sensitivity* y *Specificity*, se observa lo que ha sucedido con los anteriores algoritmos, el algoritmo clasifica correctamente los ocupados, no así los parados.

6.3.6. Comparación de los algoritmos de clasificación

Se muestran los resultados experimentales obtenidos a partir de la aplicación de los métodos de clasificación sobre las colecciones disponibles. Se observa en la Figura 6.3 la curva ROC de los distintos métodos para comparar el AUC. En la Figura 6.4 se visualiza una comparación de los distintos métodos de clasificación supervisada.

Se comparan los resultados de las distintas métricas, tomando especial atención al resultado del AUC balanceado.

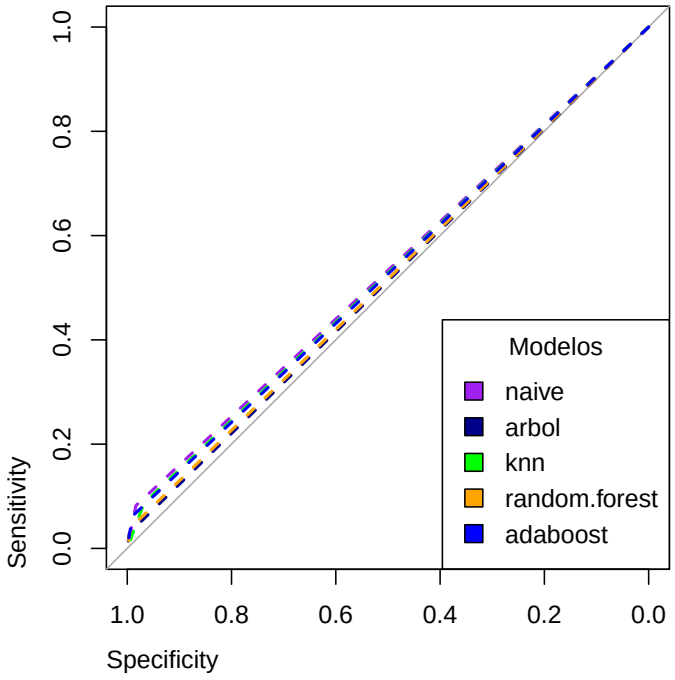
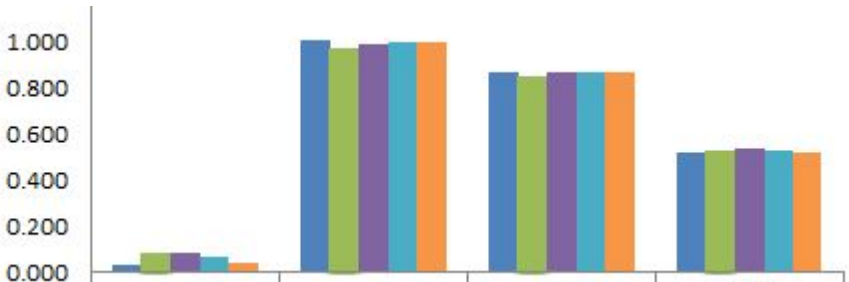


Figura 6.3: Curvas ROC de los modelos analizados.

Al analizar los resultados de forma agregada enfocándonos en el *accuracy balanceado*, que toman en consideración la correcta clasificación en ambas clases, los mejores resultados se obtienen con el modelo *naive*.

Por otro lado, si se verifica la precisión de la especificidad y la sensibilidad y de esta forma conocer el porcentaje de acierto de una clase u otra, es con *arbol* en el primero y con *k-nn* y en el segundo, con los que



	Sensitivity	Specificity	Accuracy	Accuracy balanceado
Arbol	0.030	0.996	0.864	0.513
KNN	0.087	0.968	0.848	0.527
Naïve	0.082	0.983	0.860	0.533
Adaboost	0.062	0.990	0.863	0.526
Random Forest	0.040	0.993	0.862	0.517

Figura 6.4: Comparación de modelos.

se obtienen mejores resultados. La efectividad de los indicadores de especificidad y la sensibilidad se puede visualizar mejor en la Figura 6.4. El método *arbol* es el que peor ha clasificado la clase de los parados, lo mismo que el clasificador *k-nn* para la clase ocupados. De todas formas no es que haya muchas diferencias entre un método y otro.

En general, se obtiene malos resultados para la clasificacion de parados con todos los algoritmos, ocurriendo lo contrario para la clase de ocupados.

En este capítulo se utilizaron principalmente las referencias Ripley y Venables (2020), Meyer et al. (2019), Therneau (2019), Alfaro et al. (2018), Breiman y Cutler (2018), Zhao (2015), Toomey (2014), James et al. (2013), Kuhn y Johnson (2013), Lantz (2013), Williams (2011), Maimon y Rokach (2010), Tan et al. (2006), Liaw (2002).

Capítulo 7

Conclusiones

En este trabajo se presenta la situación del mercado laboral de españa por medio del análisis exploratorio, modelos de regresión y minería de datos utilizando los datos recogidos de la Encuesta cuatrienal de Estructura Salarial de 2014 y la Encuesta de Población Activa del tercer trimestre de 2019.

Se analiza la situación en que se encuentra la persona en el mercado laboral en función de variables socio-demográficas y de estudios. Además, se realiza un estudio de los salarios para analizar cómo influyen ciertas características de las personas cruzadas con la variable `Sexo` y de esta forma conocer los determinantes de la brecha salarial entre hombres y mujeres.

Como principales resultados cabe destacar:

- En general, las variables: `Sexo, Nivel de estudio, Estado civil, Comunidad autónoma, Grupos de edad` estan relacionadas con la situación de empleo o desempleo de la persona. Esto demostrado en los análisis exploratorios, resultados del *test* de `chi cuadrado` y la regresión logística aplicada.
    - De forma particular, al contrastar los resultados del estadístico de asociación podemos concluir que de todas las variables son : `Grupos de edad, Nivel de estudio y Comunidad autónoma` las que podrían tener una mayor relación sobre los niveles ocupados o parados. Esto se confirma en el análisis de clasificación supervisada(*Random Forest y Adaboost*) al mostrar las variables más importantes a la hora de predecir si un individuo se encuentra parado o no.
- Al observar cómo influyen las variables estudiadas sobre encontrarse ocupado o parado tenemos que:
    - En cuanto a la variable `Sexo`, a pesar de que se ha avanzado de forma positiva en materia de igualdad de género en el mercado laboral, las mujeres tienen una mayor probabilidad de encontrarse en la condición de parado frente a los hombres.
    - Con respecto a los `Grupos de edad`, encontrarse en grupos de edades mayores disminuye la probabilidad de estar parado.
    - Con respecto al `Estado civil`, la probabilidad de estar parado disminuye al pasar de soltero a cualquier otro estado civil.
    - Sobre el `Nivel de estudio`, pasar del nivel de formación más bajo (analfabeto, que es el grupo de referencia) a los niveles más altos disminuye la probabilidad de estar sin trabajo.
    - El ser extranjero o tener doble nacionalidad aumenta la probabilidad de estar parado frente a ser español.
    - Con respecto a la `Comunidad autónoma` tomamos como referencia a Madrid. En los únicos casos en que la probabilidad de estar parados no aumenta y que son categorías significativas es cuando se trata de ciudades como: Aragón, Islas baleares y Navarra. Algunos casos como: Cataluña, País Vasco y Cantabria, disminuye la probabilidad pero sus p-valores están por debajo del nivel de

significación del 5 %, que entendemos que entre su clasificación de parados y ocupados no hay muchas diferencias. En los demás aumenta la probabilidad, especialmente para Melilla y Ceuta.

- En el análisis de lo salarios se concluye que existe una brecha salarial de la ganacia por hora entre hombres y mujeres. Visto que la importancia del `Sexo` en la determinación de los salarios es clara, en determinados perfiles de trabajadores, el salario es más dependiente de esta variable que en otros. Hay mujeres que por susbcaracterísticas personales y laborales, están más discriminadas que otras como se observo en el análisis:

  - Es significativa la brecha salarial entre hombres y mujeres cuanto mayor es la `Edad` de los trabajadores
  - Igualmente, mientras mayor es la `Antigüedad` en la empresa, se incrementa la diferencia de la ganancia por hora entre hombres y mujeres.
  - En los `Niveles de estudio` más altos aumenta la discriminación salarial.
  - De acuerdo a los resultados de la variable `Tipo de jornada`, las mujeres con jornada parcial estarían menos discriminadas que las mujeres que trabajan a tiempo completo.
  - Cuando los `Contratos` son de tiempo indefinido frente a contratos de duración determinada aumenta la brecha salarial significativamente.
  - El tener `Responsabilidades de supervisión` frente a no tenerlas aumenta la brecha salarial entre hombres y mujeres.
  - Las brechas más importantes de la ganancia media por hora entre hombres y mujeres de las distintas `ocupaciones` se encuentra entre las `Ocupaciones` de: “Directores y gerentes”, “Técnicos y profesionales científicos e intelectuales”, “Trabajadores cualificados de las industrias manufactureras” y “operadores de instalaciones y máquinas”.

- En general, con el análisis de minería de datos, podemos concluir que a pesar de que las variables como: `Sexo, Nivel de estudio, Estado civil, Comunidad autonoma, Grupos de edad`, influyen en la categoría de parados u ocupados, no es posible predecir si un individuo elegido al azar es parado o no a partir de estas variables.

  El hecho de que no se obtenga mejor acierto parece indicar que los factores o variables explicativas, aun siendo significativas las diferencias observadas, no son decisivos, y por lo tanto deben existir otras variables con mayor influencia para encontrar empleo.

Bibliografía

Agresti A (2007) An introduction to categorical data analysis. Wiley, New York.

Alfaro E, Gamez M, Garcia, N.(2018). adabag: Applies Multiclass AdaBoost.M1, SAMME and Bagging. R package version 4.2.https://CRAN.R-project.org/package=adabag

Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. https://CRAN.R-project.org/package=rmarkdown.

Anghel B, Conde-Ruiz JI, Marra-De Artíñano I (2018) Brechas Salariales de Género en España. FEDEA.

Bendixen M (2003) A Practical Guide to the Use of Correspondence Analysis in Marketing Research. Marketing Research On-Line 1:16-38.

Breiman, L. (1996): "Bagging predictors". Machine Learning, Vol 24,2, pp.123-140.

Breiman L, Cutler A (2018). class: randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.6-14. https://CRAN.R-project.org/package=randomForest.

CEOE (2019) Análisis de la brecha salarial de género en España. PricewaterhouseCoopers Asesores de Negocios, S.L.

Crawley, MJ (2007) The R book. Jonhn Wiley y Sons, England.

De Cabo G y Garzón, MJ (2007) "Diferencia y discriminación salarial por razón de sexo." Estudios del Instituto de la Mujer. CEET.

De Cabo G y Rodríguez M (2014) Investigación conducente a la elaboración de un índice sintético de discriminación salarial. CEET.

De Cabo G, Rodríguez M, Segales M (2013) Medición de la brecha y la discriminación Salarial en la Comunidad de Madrid. CEET, España.

Hastie T, Tibshirani R, Friedman J (2013). The Elements of Statistical Learning: Data Mining,Inference, and Prediction. Springer Series in Statistics, California.

Hosmer, D.W. y Lemeshow, S. (1989). Applied logistic regression. Wiley, USA.

Instituto Nacional de Estadística, INE (2019) Notas de prensa Encuesta de Población Activa (EPA): Tercer trimestre de 2019. INE, España.

Instituto Nacional de Estadística, INE (2017) Encuesta de Población Activa (EPA), Metodología 2005. Descripción general de la encuesta. Madrid.

Instituto Nacional de Estadística, INE (2017) Estructura Salarial (EES), Metodología. Madrid.

Instituto Nacional de Estadística, INE (2016) Notas de prensa Encuesta de Estructura Salarial: Resultados definitivos. INE, España.

James, G, Witten, D, Hastie, T, Tibshirani, R (2013) An Introduction to Statistical Learning with Applications in R. Springer.

Kuhn M, Johnson K (2013) Applied Predictive Modeling. Springer, New York.

Lantz, B (2013) Machine Learning with R. Packt Publishing Ltd, UK.

Liaw A, Wiener M (2002) Classication and Regression by randomForest. R News, 2(3), 18–22.

Maimon O, Rokach L(2010) Data Mining and Knowledge Discovery Handbook. Springer, New York.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang C, Lin C (2019) e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R packageversion 1.7-3. https://CRAN.R-project.org/package=e1071.

Montgomery, DC (2013) Design and Analysis of Experiments. J. Wiley and Sons, USA.

Quinlan, JR(1996) Bagging, boosting, and C4.5. In: Proceedings of the thirteenth national conference on artificial intelligence and the eighth innovative applications of artificial intelligence conference. AAAI Press/MIT Press, Menlo Park, pp 725–730.

Observatorio de las ocupaciones del SEPE (2019) Informe del mercado de trabajo estatal, datos 2018. Edición realizada por el Servicio Público de Empleo Estatal, Madrid. pp 18-31

Ripley B, Venables W (2020). class: Functions for Classification. R package version 7.3-17. https://CRAN.R-project.org/package=class.

Sheather, S.J. (2009). A modern approach to regression with R. Springer, New York.

Tan PN, Steinback M, Kumar V (2006) Introduction to Data Mining. Pearson Addison-Wesley, USA.

Therneau T, Atkinson B, Ripley B (2019) rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart

Toomey Dan (2014) R for Data Science. Packt Publishing Ltd, UK.

Viswanathan V, Viswanathan S, Gohil A, Yu-Wei C (2016). R:Recipes for Analysis, Visualization and Machine Learning. Packt Publishing.

Williams G (2011) Data mining with rattle and R. Springer, New York.

Zhao Y (2015) R and Data mining: Examples and Case Studies. Published by Elsevier.