MODELADO DE LA CAUSALIDAD DE LA SOBREVIVENCIA EN LOS EMPRENDIMIENTOS FORMALES EN URUGUAY

Ec. Felix Bellomo
Ec. Martín Naranja
Ec. Micaela Antúnez
Lic.T.S. Inés Núñez

La Asesoría General en Seguridad Social (AGSS) del Banco de Previsión Social (BPS) en coordinación con la Universidad ORT Uruguay – Escuela de Postgrados de Educación Ejecutiva, desarrolla un proyecto de investigación con el equipo técnico de la Asesoría en Políticas de Seguridad Social (APSS) con foco en los emprendimientos formales registrados en BPS.

Como resultado del intercambio profesional se produce este documento.

El asesoramiento de Universidad ORT para el proyecto fue brindado por los docentes Damián Coltzau y Daniel Gramoso.[4]

[4] Damián Coltzau, MSc in Statistics, Birkbeck College, University of London, Reino Unido. Máster en Economía, Universidad Torcuato di Tella, Argentina. Licenciado en Economía, Universidad de Buenos Aires. Ex investigador y consultor, Quantitative Risk Management, Londres, Reino Unido.

Daniel Gramoso, Licenciado en Economía, Universidad de la República (Uruguay). Asesor técnico, Buró de Radios del Uruguay. Director, Mediciones y Mercado. Exgerente, área de Tecnología y Consultoría Analítica, Equifax Uruguay - Clearing de Informes. Exasesor, área Estadística, Markets Panels. Coordinador académico adjunto del Diploma de Especialización en Analítica de Negocios, Facultad de Administración y Ciencias Sociales, Universidad ORT Uruguay.

Modelado de la causalidad de la sobrevivencia en los emprendimientos formales en Uruguay

Ec. Felix Bellomo, Ec. Martín Naranja, Ec. Micaela Antúnez, Lic.T.S. Inés Núñez
Asesoría en Políticas de Seguridad Social
Agosto 2022

Resumen

En este informe se estudia la sobrevivencia de los emprendimientos formales registrados en BPS que iniciaron en el período 2010-2015. Aplicando modelos econométricos y con datos administrativos del BPS se analizan cuáles son los aspectos significativos que inciden en la vida y permanencia de los mismos.
Los resultados indican que las principales características que influyen en la duración de los emprendimientos son: el tipo de actividad a la que se dedica la organización (Giro), la Naturaleza jurídica, la propia experiencia del emprendedor como tal y si la organización cuenta con trabajadores o no.

Palabras Claves

Emprendimientos - Empresas - Emprendedores - Sobrevivencia

1. Introducción

Hacia principios del 2020, entre las propuestas proyectadas por la APSS dentro de la agenda para ese año, se aprueba la investigación sobre el Emprendedurismo Senior. El proyecto implicaba atender varios desafíos, entre ellos el acopio de conocimiento orientado al estudio de diversas dimensiones como el Emprendedurismo y, desde este ámbito, revelar como se integran los Seniors (las personas de 50 y más años) a esta actividad. Se realizaron entrevistas a referentes nacionales que estudiaron este fenómeno, acumulando bibliografía internacional y aprendiendo el proceso recorrido por Uruguay desde una transformación hacia una economía integral y social.

Se complementó el estudio considerando un nuevo paradigma que se inicia desde la Economía Plateada[1] por el Banco Interamericano de Desarrollo –BID- promoviendo en nuestro país el proyecto "Pensar en Grande"[2]. Se analizó, desde la perspectiva de la seguridad social, la normativa que habilita u obstaculiza emprender entre las personas mayores -como los aspectos considerados por la Comisión de Expertos de Seguridad Social- y también si es viable articular esta modalidad para enfrentar el envejecimiento y las erogaciones de las prestaciones[3].

[1] La Economía Plateada se entiende como la parte de la economía global vinculada al cambio demográfico producido por el envejecimiento de la población cuyo enfoque se centra en las necesidades y demandas de los adultos mayores.
[2] Pensar en Grande es una iniciativa del BID, Endeavor Uruguay y Xeniors que tiene como objetivo involucrar al ecosistema de innovación, para generar un proceso virtuoso de creación de valor económico para las personas mayores con énfasis en aquellas más vulnerables.
[3] https://www.bps.gub.uy/bps/file/18022/1/emprendedurismo-senior-en-uruguay-el-envejecimientocomonueva-oportunidad-para-crecimiento.-m.antunez-m.naranja-y-i.nunez.pdf

La construcción de una base de datos implicó la conjunción de fuentes de información provenientes de distintas oficinas del BPS, así como desde el inicio crear un sistema de datos que ofreciera las garantías para el análisis[4].

Dicha investigación logró los objetivos previstos constituyendo el marco de análisis del fenómeno a nivel internacional y su aplicación en el ámbito nacional, complementando un estudio cuantitativo en base a datos administrativos, exhibiendo el comportamiento de los emprendimientos registrados formalmente en BPS y evidenciando una evolución positiva de los seniors.

Durante los últimos años, el fenómeno del Emprendedurismo tomó impulso haciendo foco esencialmente en los jóvenes emprendedores, pero es el envejecimiento poblacional el que ahora invita a un cambio de paradigma y fija grandes expectativas sobre el envejecimiento activo.

En el marco del desarrollo de la Economía Plateada en América Latina y el Caribe, la propuesta de investigación se relaciona con el fenómeno del emprendedurismo senior en el Uruguay.

Las personas mayores cuentan con un gran potencial para emprender, bajo el entendido de que la innovación y la creatividad se logran potenciar con la acumulación de experiencia en la trayectoria laboral y personal.

A su vez, es esperable que el estímulo a las capacidades de este colectivo pueda colaborar tanto con el empoderamiento como con el bienestar personal, revalorando su participación social y económica a través de nuevas actividades que dinamicen la inversión, generen nuevas fuentes laborales, promuevan la capacitación y la reconversión en el marco de un envejecimiento activo y satisfactorio.

Dimensionar el emprendedurismo en nuestro país, caracterizándolo e identificando sus principales desafíos, en el marco de su vinculación con la seguridad social y dentro del contexto de la discusión de la reforma de seguridad social en Uruguay, implica la profundización de estudios que demuestren que aspectos son relevantes al momento de aplicar una política de estímulo o fomento.

Contar con un análisis descriptivo del emprendedurismo, su proyección y el estudio de los factores que tienen una mayor incidencia en relación a la sobrevida y el impacto que generan sobre el empleo, permitiría brindar elementos sustantivos a los hacedores de políticas públicas para la toma de decisiones.

En base a los criterios establecidos por BID, GEM y OCDE, se inicia el estudio en relación a la sobrevivencia de los emprendimientos como medida de éxito, concretamente en aquellos que alcanzan una duración en el tiempo de 60 meses o más (desde la fecha de su inicio). Una variable de especial significación para este estudio es la Edad del Emprendedor, en relación al objeto de la investigación de los emprendedores seniors (50y+).

[4] https://www.bps.gub.uy/bps/file/18734/1/82.-emprendedurismo-senior-en-uruguay.-caracter.-y-analisis-de-los-empredededores-afiliados-a-bps.-nunez-naranja-y-antunez.pdf

Se pretende estudiar la significación de diversos factores en el éxito emprendedor, distinguiendo las características que hacen a la sobrevida de los emprendimientos, para poder contrastar los resultados obtenidos con las conclusiones que surgen a nivel internacional.

Habiendo obtenido un desarrollo importante sobre la temática del Emprendedurismo, la gerencia de la AGSS aprueba la propuesta de contar con una tutoría de manera de disponer de una formación que permita determinar los aspectos que inciden en la dinámica empresarial formal registrada en BPS.

2. Conceptos principales

Emprendedor y Emprendimiento _

GEM[5] define como "nuevos emprendedores a aquellos que han pagado más de 3, pero menos de 42 meses de sueldos"

A partir de ello, proponemos clasificar como "emprendedores formales" a los no dependientes registrados en BPS a cargo de "emprendimientos", entendidos estos como las empresas activas que hayan presentado más de 3 pero menos de 42 nóminas.

Empresa _

Una vez que un "emprendimiento ha pagado sueldos por más de 42 meses", deja de ser considerado un emprendimiento, pasando a ser una "empresa" según GEM.

Es por ello que proponemos identificar como "empresas" a los emprendimientos con más de 42 nóminas presentadas.

Sobrevivencia _

Se considera el tiempo de vida de las empresas o su permanencia dentro del ciclo económico.

Éxito emprendedor _

Si bien para GEM el éxito emprendedor se puede definir de diversas maneras (consolidándose como empresa, de acuerdo a la antigüedad, alcanzando el status de Empresa de Rápido Crecimiento, a la cantidad de puestos alcanzados, etc.), para nuestro análisis el éxito vendrá representado para aquellos emprendimientos que tengan una sobrevida mayor o igual a los 60 meses.

Según el marco teórico, alcanzar el éxito depende de diversos factores que pretendemos validar a través del estudio de las siguientes dimensiones:

[5] Global Entrepreneurship Monitor (GEM), es un estudio sobre el estado del emprendimiento a nivel mundial.

- Motivación para emprender (necesidad / oportunidad). Se propone analizar a partir de la información relevada acerca de la causal de baja, causal jubilatoria, si emprende en paralelo a su actividad como dependiente, etc.

- Experiencia del emprendedor en la misma rama de actividad del nuevo emprendimiento.

- Experiencia anterior como emprendedor. Se propone identificar si se ha desempeñado en el pasado como no dependiente.

3. Objetivo

Nuestro estudio se centrará en poder identificar los principales aspectos que inciden en una mayor sobrevivencia de los emprendimientos por medio de diversas herramientas para el análisis de datos.

De forma complementaria, capacitar al equipo de investigación en el manejo de modelos e ingeniería de la información con el programa estadístico R.

4. Metodología aplicada

En este estudio se sigue la metodología CRISP-DM (Cross Industry Standard for Data Mining) que, como modelo de proceso, ofrece un resumen del ciclo vital de un proyecto de minería de datos.

Figura 1 _ Fases del proceso de CRISP-DM

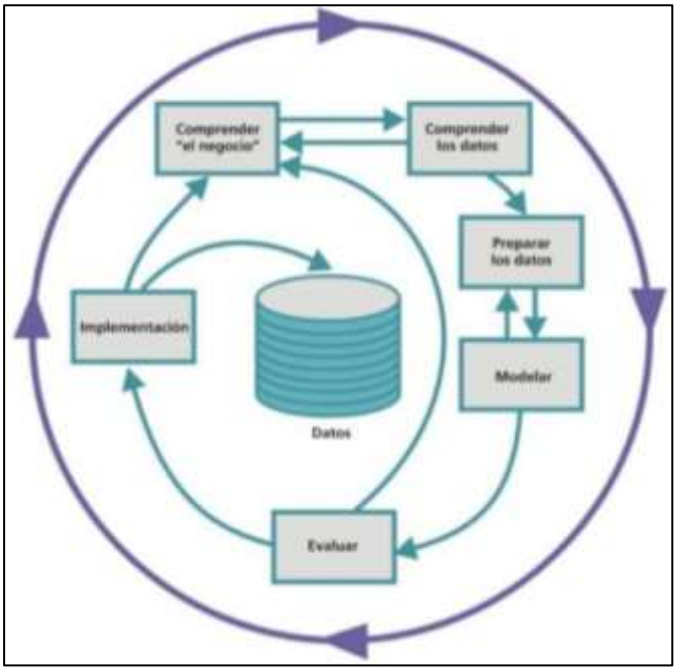


Imagen de web.[6]

[6] en https://www.researchgate.net/figure/Fases-del-proceso-de-CRISP-DM-Adaptado-de-10_fig2_306959832

El ciclo de vida del modelo presenta seis fases con flechas que indican las dependencias más importantes y frecuentes entre las fases (pero la secuencia de las fases no es rígida). El círculo externo en la figura simboliza la naturaleza cíclica de los proyectos de análisis de datos.

La fase inicial se debe enfocar en la comprensión de los objetivos del proyecto, definiendo el problema (minería de datos), diseñando un plan para alcanzar los objetivos. En la fase de entendimiento de datos se deben identificar los problemas de calidad, así como ampliar el conocimiento preliminar de los mismos. En la preparación de los datos se construye el conjunto final de datos a utilizar en las herramientas de modelado. En la fase de modelado se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema. En la siguiente etapa se debe evaluar el modelo (o los modelos) alcanzado, verificando que se haya logrado un cierto estándar de calidad. Finalmente se despliegan los resultados de manera que puedan ser utilizados por los interesados.

4.1 – Comprendiendo el negocio

Teniendo en cuenta el objeto y las hipótesis del estudio se procedió a confeccionar la base de información necesaria para realizar el análisis previsto, acordando a medida que fue transcurriendo la tutoría las mejores herramientas de utilización mediante, como ser una comparación de técnicas entre árboles de clasificación y regresiones, ampliando a su vez los conocimientos en el manejo analítico del software R.[7]

El análisis se centró en la dinámica empresarial, diagnosticando los aspectos relevantes que muestran una mayor incidencia en la sobrevida de los emprendimientos formales inscriptos en BPS, distinguiendo además los casos en que sus patrones tienen trabajadores dependientes en su plantilla, o no tienen.

4.2 – Preparación y entendimiento de los datos

Para realizar el análisis de los emprendimientos se confeccionó una base de datos en la que se unió la información de los datos de Registro de Empresas (el tipo de empresa, la fecha de inicio, su localización, código CIIU, etc.), de Puestos de Trabajo (cantidad de puestos, remuneraciones, el vínculo funcional, etc.), de una muestra de la Historia Laboral[8] (conteniendo información de aproximadamente 150 mil personas, de las que se utilizaron finalmente 8513 casos de emprendedores) y de un conjunto de variables macroeconómicas (el PBI a pesos constantes, la tasa de desempleo, etc.), para el período de agosto 2006 a diciembre 2020.

Con el objetivo de estudiar la sobrevida de las empresas en el mediano plazo, considerando que el tiempo necesario para lograr el éxito requiere de por lo menos 60 meses, se propone dar seguimiento a los emprendimientos registrados desde 2010 a 2015, ya que al momento de realizar la tutoría se contaba con la información completa hasta el año 2020.

[7] A partir de un cronograma de actividades definidas entre los docentes de ORT y el equipo técnico del BPS, se acordó un esquema de trabajo que incluyó clases virtuales y presenciales durante octubre/2021 a enero/2022.
[8] Esta información surge de una base auxiliar de BPS que cuenta con una muestra de personas con su respectiva historia laboral durante el período abril de 1997 a diciembre de 2019. Para estos 8513 registros se calculó un expansor (ponderador) que representa el total de 131.116 registros de emprendimientos. Este ponderador se construyó tomando en cuenta la distribución de los emprendimientos totales de acuerdo a la naturaleza jurídica

Inicialmente la base original tiene 94 variables detalladas en el Anexo “A”. Luego en el proceso se hace una selección de las variables más importantes. Para ello se realizaron los correspondientes análisis univariado y de correlación bivariada de las variables independientes de forma de conseguir una mejor comprensión de la distribución de las variables a través de diferentes tipos de gráficos y lectura de los datos estadísticos.

Se eliminaron los giros B, D, E, O, U[9], por considerarlas actividades con características muy particulares (trabajadores en embajadas, asesores en organizaciones internacionales, etc.), además de que para cada una de ellas la cantidad de casos es muy baja, poniendo en riesgo la representatividad en cuestión de los datos.

La presentación de los datos tiene un enfoque de sección transversal, si bien los emprendimientos no inician simultáneamente, nuestro interés radica en observar su duración, más allá del momento del tiempo en que se concreten. [10]

4.2.1 Análisis univariado

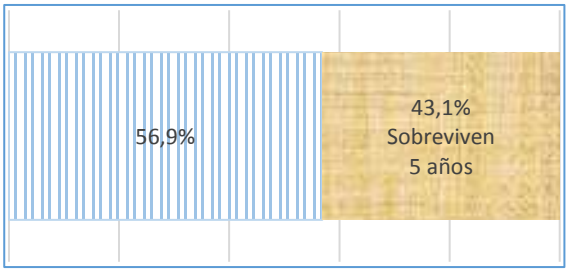
Previo a desarrollar cualquier tipo de procesamiento de los datos, se realiza el análisis estadístico univariado de todas las variables que se emplean en este estudio, de manera de comprender mejor las características de nuestros datos. A continuación se presenta el análisis para las principales variables que se tomaron en cuenta.

En lo que refiere a la supervivencia, menos de la mitad de los casos extienden su permanencia por un período mayor a los 5 años.
En cuanto al tipo de aportación a la Seguridad social, la distribución de los casos en la base es semejante a la del total de los casos en los registros administrativos, donde la aportación de Industria y Comercio es la amplia mayoría.

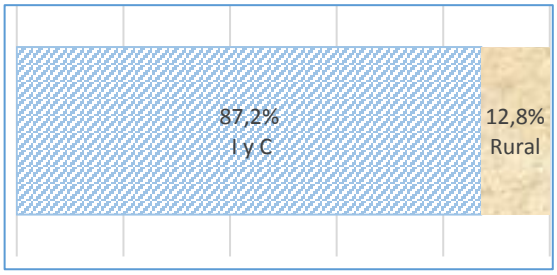
[9] B: Explotación de minas y canteras; D: Suministro de electricidad, gas, vapor y aire acondicionado; E: Suministro de agua, alcantarillado, gestión de desechos y actividades de saneamiento; O: Administración pública y defensa, planes de seguridad social de afiliación obligatoria; U: Actividades de organizaciones y órganos extraterritoriales.
[10] Para Croissant, esta manera de exponer los datos “es peculiar respecto a contextos experimentales, ya que enfatiza la especificación y prueba del modelo y aborda un número de cuestiones que surgen de los problemas estadísticos particulares asociados con la economía de los datos”. Yves Croissant, Giovanni Millo; Econometría de Datos de Panel en R: el paquete plm; Diario de Sta Software estadístico; Universidad de California, Los Ángeles, 2008.

Gráfico 1_

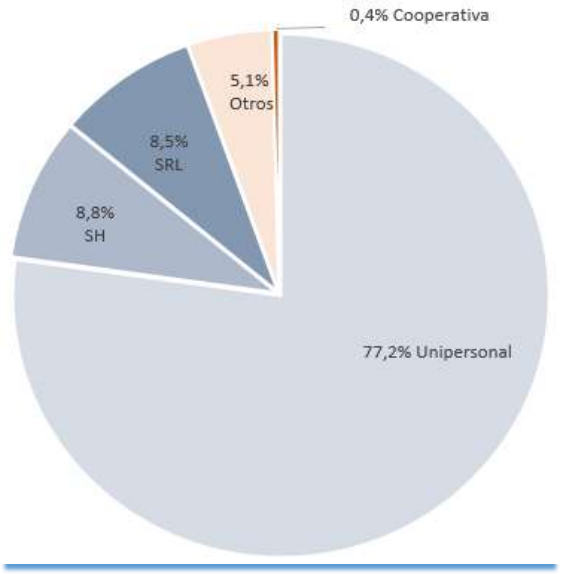
A_ proporción de sobrevivencia T5



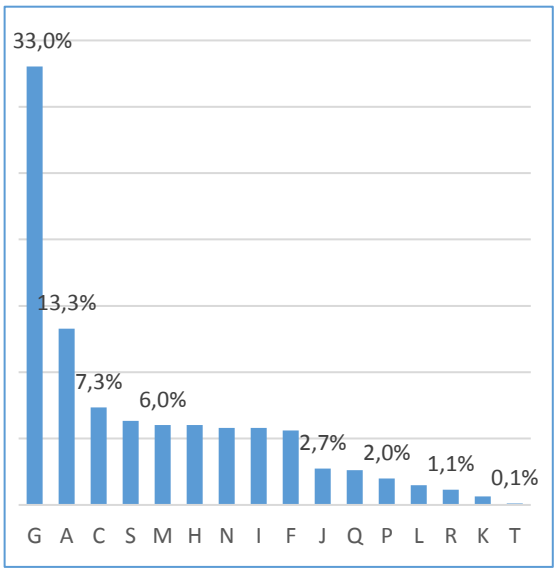
B_ distribución por tipo de aportación



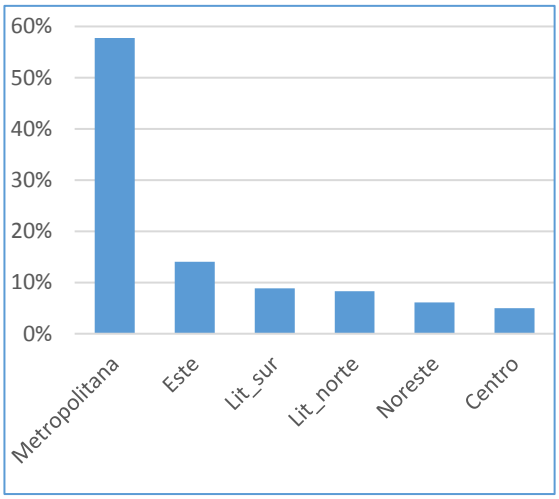
C_ naturaleza jurídica [11]



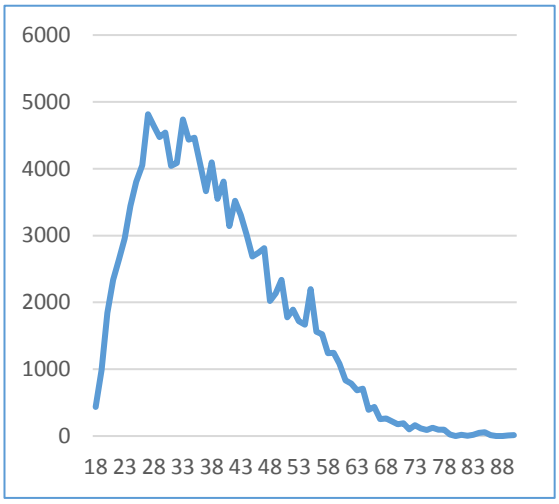
D_ giro de actividad [12]



E_ zona geográfica [13]



F_ edad de los emprendedores



[11] SH- Sociedad de hecho; SRL- Sociedad de responsabilidad limitada; Otros- incluye SA

[12] A- Producción agropecuaria, forestación y pesca; C -Industrias manufactureras; F - Construcción; G - Comercio al por mayor y al por menor, reparación de vehículos de motor y de las motocicletas; H - Transporte y almacenamiento; I -Alojamiento y servicios de comida; J - Información y comunicación; K- Actividades financieras y de seguros; L - Actividades inmobiliarias; M- Actividades profesionales, científicas y técnicas; N- Actividades administrativas y de servicios de apoyo; P - Enseñanza; Q - Servicios sociales y relacionados con la salud humana; R -Arte, entretenimiento y recreación; S - Otras actividades de servicio; T - Actividades de los hogares en calidad de empleadores, etc..

[13] **Centro**: Durazno, Flores y Florida; **Este**: Treinta y tres, Lavalleja, Maldonado, Rocha; **Litoral_norte**: Paysandú, Artigas y Salto; **Litoral_sur**: Colonia, Río negro y Soriano; **Metropolitana**: Canelones, Montevideo y San José; por último **Noreste**: Cerro largo, Rivera y Tacuarembó

Del mismo modo, no es de extrañar que más de 3 de cada 4 casos correspondan a emprendimientos del tipo Unipersonal, así como que un tercio de los emprendimientos integren la rama de actividad G - "Comercio al por mayor y al por menor; reparación de los vehículos de motor y de las motocicletas".
En lo que refiere a la distribución territorial la concentración es notoria en la zona metropolitana, y también en el Este se observa un peso importante de los emprendimientos.

En cuanto a características más vinculadas a los emprendedores, encontramos que la mayoría son hombres (en una relación similar a la de los cotizantes totales a la Seguridad social en el período), y que dentro de los emprendedores la mitad inicia experiencias antes de los 36 años aproximadamente.

En general los motivos que se tienen para emprender pueden agruparse en "oportunidad" y "necesidad", donde la oportunidad refiere a un nicho de mercado sin explotar o a una idea original (entre otras), mientras la necesidad se vincula con el hecho de trabajar para generar ingresos fundamentalmente. Algunas variables nos pueden dar pistas sobre el motivo que el emprendedor tiene para iniciar un emprendimiento.
En el año previo a iniciar los emprendimientos, 1 de cada 3 emprendedores presenta un egreso voluntario (renuncia) a alguna actividad que lo encontraba como dependiente, en tanto quienes fueron despedidos o finalizaron su contrato representaron a 1 de cada 7.
Aproximadamente 1 de cada 10 presentó un período con subsidio de desempleo y 1 de cada 9 fue beneficiario de subsidio por enfermedad. Apenas 2 de cada 100 fueron beneficiarios de algún subsidio por maternidad o paternidad en el año previo.

Cuadro 1 _ análisis univariado

Variables	unidad	medida	#	medida	#
Sexo	(%)	hombres	55,7	mujeres	44,3
Edad inicio	(años)	mediana	36,0	media	38,4
egreso voluntario	(%)	Sí	34,6	No	65,4
egreso forzado	(%)	Sí	14,7	No	85,3
con desempleo	(%)	Sí	10,4	No	89,6
con enfermedad	(%)	Sí	11,2	No	88,8
con mater-pater	(%)	Sí	2,2	No	97,8
Q_desempleo	(meses)	mediana	0,0	media*	2,8
fue patrón	(%)	Sí	22,2	No	77,8
Q_fue patrón	(meses)	mediana	0,0	media**	19,4
No_depte_t0	(%)	Sí	13,8	No	86,2
Patr_sDep	(%)	Sí	67,6	No	32,4
Mismo giro	(%)	Sí	54,7	No	45,3
Mismo cód giro	(%)	Sí	15,8	No	84,2
sin registro	(meses)	mediana	59,0	media	71,9
prom_sueldo ***	($)	mediana	7666	media	12488

Notas: *) Para los que presentan desempleo la media es de 9,9
**) para los que presentan experiencia como patrón la media es de 60,1
***) prom_sueldo se expresa a valores del año 2014

En lo que refiere a la historia de trabajo de las personas, las trayectorias laborales de éstas pueden verse alteradas por haber atravesado períodos de desempleo o de enfermedad, así como también de la experiencia (o capacitación) que puedan haber adquirido.
En media, al inicio del emprendimiento los emprendedores presentan casi 3 meses en su historia laboral con uso del subsidio por desempleo (aunque debe notarse que más de la mitad no han hecho uso del mismo).
Sólo 2 de cada 10 tienen experiencia siendo patrones, lo que genera que la media presentada sea de aproximadamente 1 año y medio.
Al momento de iniciar el emprendimiento, 1 de cada 7 emprendedores se declaraba como no dependiente en otra empresa. Estos casos pueden corresponder a situaciones en que se finaliza un emprendimiento y se inicia otro, o a que algunos emprendedores tienen más de un emprendimiento a la vez.

A su vez, en 2 de cada 3 casos los emprendedores no contratan trabajadores dependientes (por el tipo de emprendimiento, porque no los precisan o por restricciones económicas o de otro tipo), y en más de la mitad de los casos quienes emprenden lo hacen en una rama de actividad en la que anteriormente habían trabajado, aunque una minoría (3 de cada 20 aprox.) emprenden específicamente en el mismo sector (mismo código de giro).[14]

La variable sin registro se genera como aproximación para contar la cantidad de meses (potenciales) que a una persona le falta tener registrada en su historia laboral[15].
Los ingresos se expresan como un promedio para el año anterior al inicio del emprendimiento, todos llevados a precios del año 2014.

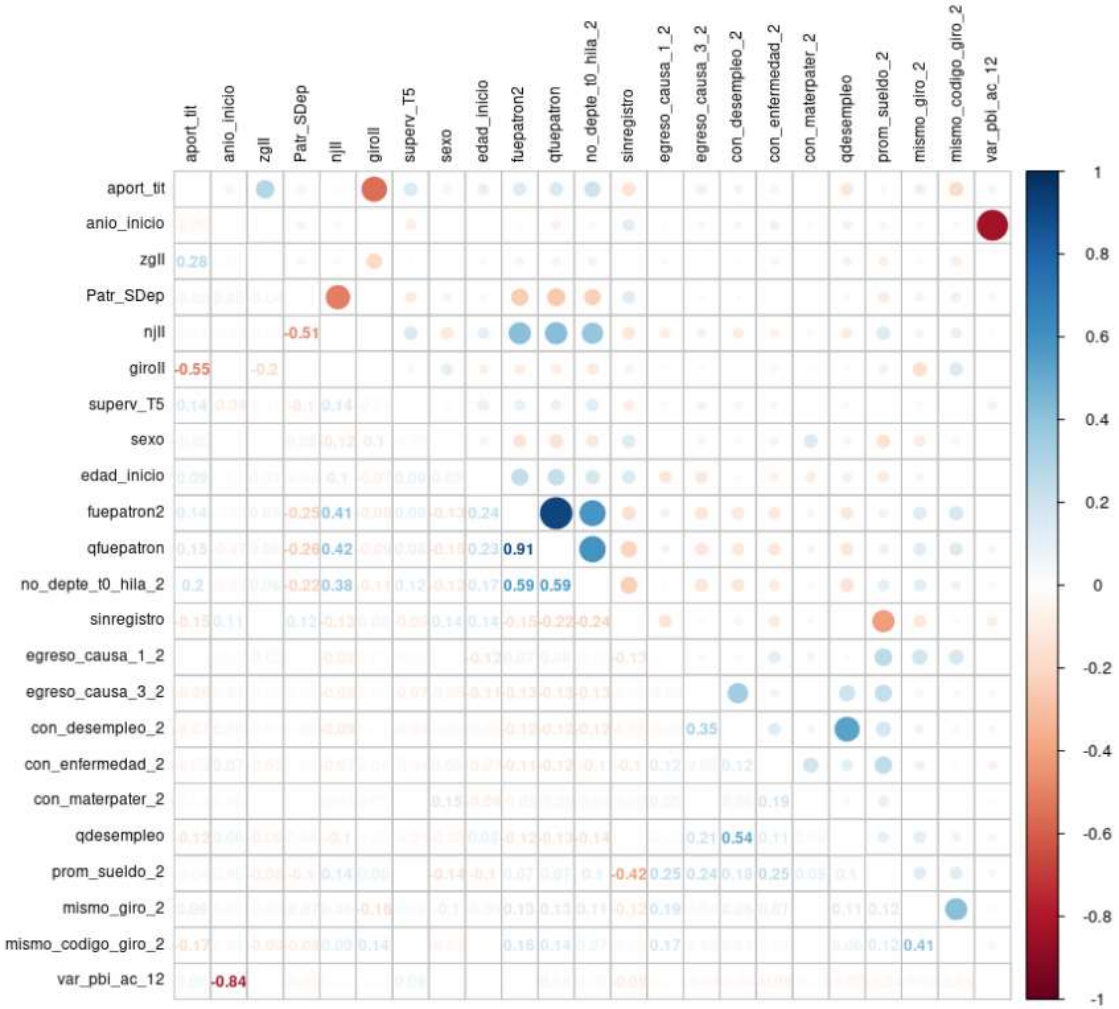
4.2.2 Análisis de correlación de las variables independientes

Para conocer las relaciones entre las variables empleadas (tomadas de dos a dos) se realiza el cálculo de la matríz de correlación bivariada por el método de Spearman. Este método mide la asociación o interdependencia entre dos variables de todas las analizadas y oscila entre -1 y 1 (asociaciones negativas, cero, asociaciones positivas). El método evalúa la relación monótona entre 2 variables continuas u ordinales. En una relación monótona las variables tienden a cambiar al mismo tiempo, pero no necesariamente a un ritmo constante.

En la siguiente gráfica se presentan los coeficientes de correlación obtenidos, mostrando las relaciones existentes entre cada variable y las restantes analizadas, donde la escala de colores, la nitidez de los números y el tamaño de los círculos indica el grado de correlación entre cada variable.

[14] Por ejemplo, la misma rama G de comercio al por mayor y al por menor, pero un giro diferente; 47522 - Comercio al por menor de pinturas y revestimientos similares, 47525 - Comercio al por menor de materiales eléctricos y conducción eléctrica. (CIIU, Revisión 4).
[15] Para su generación, se tomaron varios supuestos, por ello se entiende esta variable como una aproximación simplemente.

Gráfico 2 _ Correlación de Spearman



Nota: escala de correlación –en valor absoluto-, muy débil o inexistente (entre 0 y 0,20), débil (entre 0,21 y 0,40), moderada (entre 0,41 y 0,60), fuerte (entre 0,61 y 0,80) y muy fuerte (entre 0,81 y 1).

Puede visualizarse entonces que las variables que presentan las correlaciones más fuertes son “fuepatron2 - qfuepatron” (si fue patrón y la cantidad de meses que lo fue) con una asociación positiva y “anio_inicio - var_pbi_ac_12” (el año en que inicia el emprendimiento y la variación acumulada del PIB) con una asociación negativa. Para el segundo caso, en el periodo de estudio se evidenció una disminución del dinamismo en la economía local.

Luego, para una correlación moderada encontramos “no_depte_t0_hila_2 – fuepatron2” y “no_depte_t0_hila_2 – qfuepatron” (si el emprendedor se encontraba al momento de iniciar el emprendimiento como patrón en otro, en relación a si fue patrón, y a la cantidad de meses que fue patrón, respectivamente) con una asociación positiva. También “qdesempleo – con_desempleo2” (en relación a si tuvo desempleo en el período anterior, y a la cantidad de meses con desempleo en su historia laboral), y de manera análoga “qfuepatrón – fuepatrón2” y “mismo_giro – mismo_código_giro”.

Con una asociación negativa (y moderada) encontramos “Patr_SDep – njll” (si el emprendedor es un patrón sin dependientes y la naturaleza jurídica de su emprendimiento) y “giroll – aport_tit” (la rama de actividad del emprendimiento y el tipo de aportación). Estos casos comentados de asociación negativa no tienen a priori una explicación evidente y probablemente se deba a la forma en que fueron codificadas las variables.

Análogamente –aunque para una asociación positiva -se observa un comportamiento similar entre la naturaleza jurídica y si fue patrón (y la cantidad de meses).

También como es de esperar existe una asociación negativa y moderada, entre “sinregistro – prom_sueldo”.

En resumen, la mayoría de las variables no muestran una relación fuerte con ninguna otra, es decir, no presentan correlaciones de mayor importancia, lo que significa que no se evidencian problemas importantes de correlación.[16]

4.3 – Modelos

Emplearemos dos técnicas para identificar las variables que presentan mayor influencia al logro de una mayor supervivencia.

Por un lado presentaremos el análisis basado en árboles de decisión, que son uno de los algoritmos más utilizados para la toma de decisiones en “machine learning”, siendo fáciles de implementar y sencillos de interpretar[17].

Por otro, presentaremos el análisis basado en un modelo de regresión que buscará determinar la relación entre una variable dependiente, con respecto a otras variables, llamadas explicativas o independientes. Asimismo, el modelo buscará determinar cuál será el impacto sobre la variable dependiente ante un cambio en las variables explicativas.

4.3.1 – Árboles de decisión

Es una técnica que permite analizar decisiones secuenciales basadas en el uso de resultados y probabilidades asociadas. El árbol predice la variable dependiente con base en el aprendizaje de reglas de decisión inferidas desde las características que poseen los datos; si la variable dependiente es categórica decimos que es un árbol de clasificación, y si es numérica es un árbol de regresión (Cardona, 2009).[18]

La mayoría de los algoritmos utilizados para construir un árbol son variaciones de uno genérico llamado “Greedy algorithm”, éste busca la solución óptima en cada etapa antes de llegar al resultado final (este tipo de algoritmos son muy “codiciosos” porque van a encontrar la mejor solución de cada paso pero, en conjunto, puede que no sean la mejor solución al problema completo). La construcción del árbol sigue un enfoque de división binaria recursiva (top-down greddy approach), y la búsqueda está basada en probabilidades (se minimiza el error de clasificación, el índice de Gini y/o la Entropía. Cada vez que se hace una nueva división en el árbol, se compara el grado de impureza del nodo padre respecto del grado de impureza de los nodos hijos).

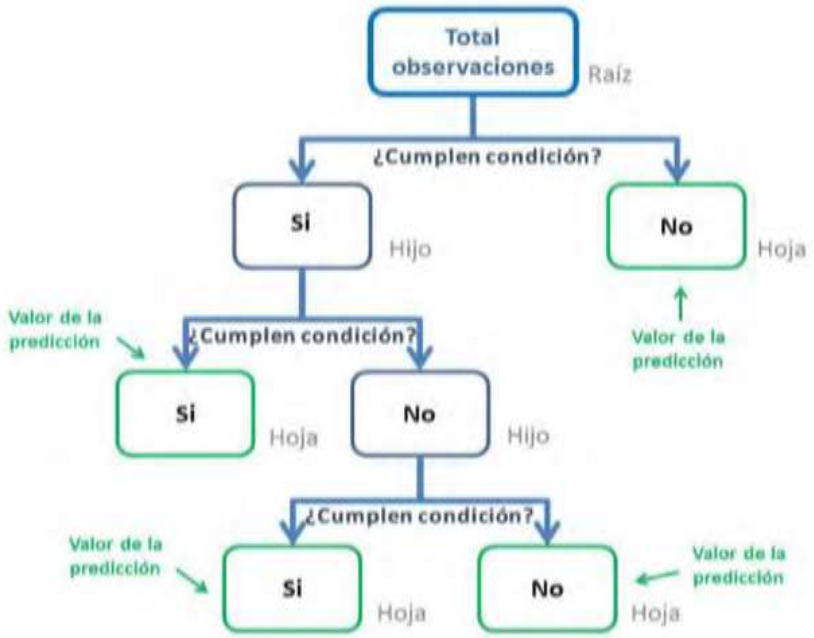
[16] Aquellas con correlación moderada, en general, responden a las variables dummy con las que distinguen la cantidad de meses que presentan dicha característica.
[17] https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/
[18] https://rpubs.com/elfenixsoy/arbol-veronica#:~:text=Los%20%C3%A1rboles%20de%20decisi%C3%B3n%20es,problemas%20(Beltr%C3%A1n%2C%202015).

Los elementos de la representación de los árboles de decisión son: raíz, nodos, ramas y hojas. El nodo raíz y los nodos internos del árbol corresponden a una prueba del valor de una de las propiedades y las ramas nodo son identificadas mediante los posibles valores de la prueba. En los nodos hoja del árbol se especifica el valor que se debe producir en el caso de alcanzar dicha hoja. El valor en el nodo terminal se reduce a la moda de las observaciones del conjunto de entrenamiento que se han encontrado en ese entramado.

Figura 2 _ Diagrama de árbol

A modo de ejemplo se presenta un diagrama explicativo de cómo se muestran los resultados del algoritmo en un árbol de decisión.

La cantidad de niveles que presente el árbol dependerá de la profundidad indicada en la programación.



Fuente: imagen web (nota 16)

Del análisis del árbol podemos observar cuáles son las variables más importantes que el algoritmo identifica para generar las divisiones en cada uno de los nodos, en definitiva para nuestro caso, las variables que el algoritmo entiende más relevantes para caracterizar a los diferentes "grupos" de emprendimientos en cuanto a lograr la sobrevida estipulada.

La técnica de aprendizaje supervisado que usaremos para obtener árboles de decisión es conocida como **CART: Classification And Regression Trees.** Se tiene una variable objetivo (dependiente)[19] y nuestra meta es obtener una función que nos permita predecir, a partir de variables predictoras (independientes), el valor de la variable objetivo para casos desconocidos.

Más información del proceso de árboles se presenta en el Anexo B.

4.3.2 – Regresión Logística[20]

La regresión logística (1958, David Cox) es un método que permite estimar la probabilidad de una variable cualitativa binaria en función de una o más variables cuantitativas y cualitativas. Permite calcular la

[19] En nuestro caso la variable dependiente es una dummy, nuestro árbol será de clasificación.
[20] REGRESIÓN LOGÍSTICA - Universidad ORT Uruguay -2020, presentación.

probabilidad de que la variable dependiente pertenezca a cada una de las dos categorías en función del valor que adquiera la o las variables independientes.

En el modelo de regresión logística, los efectos de las variables explicativas sobre la variable dependiente no son lineales[21].

Para realizar la estimación del modelo debemos recurrir al método de máxima verosimilitud. La función de verosimilitud es la multiplicación de las funciones de probabilidad para todas las observaciones que hay en la muestra. La máxima verosimilitud da la probabilidad de los ceros observados y unos en los datos.

En cuanto al modelo, la variable dependiente es una dummy que identifica a los emprendimientos como exitosos si han perdurado 60 o más meses activos (0=No sobrevive T5; 1= Sí sobrevive T5).

Las variables independientes incluyen las características del emprendimiento, las características del emprendedor (muestra de la base HILA) y alguna variable macroeconómica.

Dado que en nuestro caso la variable de interés es binaria utilizaremos un modelo Logit de la forma

$$Y = f(\beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k) + \epsilon$$

donde *f* es la función logística

$$f(z) = \frac{\exp(z)}{1 - \exp(z)}$$

Y se cumplirá que

$$E[Y] = P(Y = 1) = \frac{\exp(\beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}{1 + \exp(\beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}$$

realizándose esta estimación mediante el método de Máxima Verosimilitud.[22]

Dado que es un modelo no lineal, no resulta posible interpretar directamente las estimaciones de los parámetros β, pero sí su signo. Si es positivo, significará que un incremento en la variable asociada, causará un incremento en *P(Y=1)* -aunque desconoceremos su magnitud-, y lo contrario pasará si el estimador presenta signo negativo.

Un concepto que ayuda a profundizar en la interpretación de los estimadores es el de "Odds"[23], que se define como un cociente de probabilidades:

$$Odds = \frac{P(Y = 1)}{1 - P(Y = 1)} = \exp(\beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

[21] En el modelo de regresión lineal, β_1 se interpreta como el cambio promedio en Y asociado con un aumento de una unidad en X_1 dejando todo lo demás constante.
[22] Los estimadores de máxima verosimilitud del modelo Logit son insesgados, consistentes y eficientes.
[23] El término Odd en inglés se refiere a la razón que se establece entre la ocurrencia -o su probabilidad- de un suceso respecto a su no ocurrencia. Se interpreta como ventaja comparativa, o como razón de probabilidades.

Puede obtenerse una expresión lineal para este modelo tomando logaritmos neperianos en la reciente expresión, de donde:

$$Logit[P(Y=1)] \equiv \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_1 + \beta_2 X_2 \ + \ \ldots \ + \beta_k X_k$$

Entonces, puede apreciarse que el estimador del parámetro β_2 se podrá interpretar como la variación en el término Logit (el logaritmo neperiano del cociente de probabilidades) causada por una variación unitaria en la variable X_2 (céteris páribus).

Es así entonces que aparece el concepto de "odds-ratio", entendiéndose el mismo como el cociente entre el odds obtenido tras realizar el incremento en la variable explicativa, y el anterior al mismo. Un incremento unitario en la variable Xi genera:

$$Odds - ratio = \frac{Odds2}{Odds1} = \exp(\beta i)$$

Entonces, un coeficiente β_i cercano a 0 – equivalente a un Odds-ratio cercano a 1 – significa que cambios en la variable explicativa X_i asociada no tendrán efecto alguno sobre la variable dependiente Y.

Se pretende entonces identificar las variables que impliquen cambios en ese ratio de probabilidad, haciéndolo variar de forma significativa. Si para una determinada característica el valor de Odd ratio es mucho mayor que 1, implica que cualquier incremento en los niveles de la variable tendrá un efecto significativo sobre la variable dependiente, por lo tanto, poseer dicha característica supondría una ventaja frente a la probabilidad de ocurrencia de un evento (en nuestro caso, Sobrevivir a los 5 años).

4.3.3 – ¿Qué tan bueno será nuestro modelo?

Una forma de resumir la bondad del ajuste del modelo es la curva ROC (Relative Reciever Operating Characteristic). La curva ROC es una representación gráfica de la sensibilidad frente a (1 − especificidad) para un sistema clasificador binario según varía el umbral de clasificación. Cuando el modelo predice correctamente la totalidad de los casos, el área por debajo de la curva es igual a la unidad. En otras palabras, esto significa –para nuestro caso- que el porcentaje de "emprendimientos sobrevivientes" bien clasificados es 100% y el porcentaje de "emprendimientos que no sobreviven" mal clasificados es 0%. Un área de 0,5, es igual al resultado de un modelo que clasifica aleatoriamente los casos. Cuanto mayor es el área por debajo de la curva ROC (denominada en la literatura relacionada como AUC) mejor el modelo.[24]

A modo de guía para interpretar las curvas ROC se han establecido los siguientes intervalos para los valores de AUC (Balsa, 2017):
Valores entre 0.5 y 0.6: el modelo ajustado no es el adecuado.
Valores entre 0.6 y 0.75: el modelo ajustado tiene una tasa de clasificación regular.
Valores entre 0.75 y 0.9: el modelo ajustado tiene una buena tasa de clasificación.
Valores entre 0.9 y 0.97: el modelo ajustado tiene una tasa de clasificación muy buena.
Valores entre 0.97 y 1: el modelo ajustado tiene una excelente tasa de clasificación.

[24] La mayoría de los paquetes estadísticos pueden adaptarse a modelos de regresión logística lineal por máxima versosimilitud. En el paquete R se utilizó la función glm.

4.4 – Evaluación y resultados

Para analizar el poder predictivo de las variables por categorías respecto a la variable dependiente, es necesario utilizar el apoyo de una serie de indicadores llamados Ap_odd, weight of evidence (WOE), information value (IV) y el indicador K-S (kolmogorov-smirnov).

Ap_odd se calcula como la cantidad de los casos ausentes (no alcanzan la sobrevida a los 60 meses) sobre la cantidad de casos presentes (los que sí alcanzan). Si el cociente es menor de 1, quiere decir que hay más casos que logran sobrevivir al menos 60 meses que los que no lo hacen, mientras que si es mayor a 1 ocurre lo contrario.

WOE significa “peso de la evidencia", es una forma de codificación de la variable independiente original y se calcula de esta forma: *ln*(%no-evento/%evento). Se entiende –para nuestro caso- que a menor WOE es mayor el peso del grupo para explicar la ocurrencia de la sobrevivencia.

Por su parte el indicador IV es WOE multiplicado por la diferencia entre la proporción de respuestas y la proporción de no respuestas en este grupo. El valor de información es una de las técnicas más usadas para seleccionar las variables más importantes en un modelo predictivo. Ayuda a clasificar las variables en función de su importancia. Valores mayores a 0.3 denotan un fuerte poder predictivo.

El indicador K-S (Kolmogorov-Smirnov) también nos ayuda para ver la efectividad del modelo, cuanto mayor sea el valor del indicador más efectivo es el modelo para capturar las respuestas (valores de K-S mayores a 20 se consideran aceptables para el modelo).

En esta parte del proceso, nuestro objetivo fue determinar que variables presentan una mayor importancia para poder explicar la supervivencia mayor a 5 años de los emprendimientos.

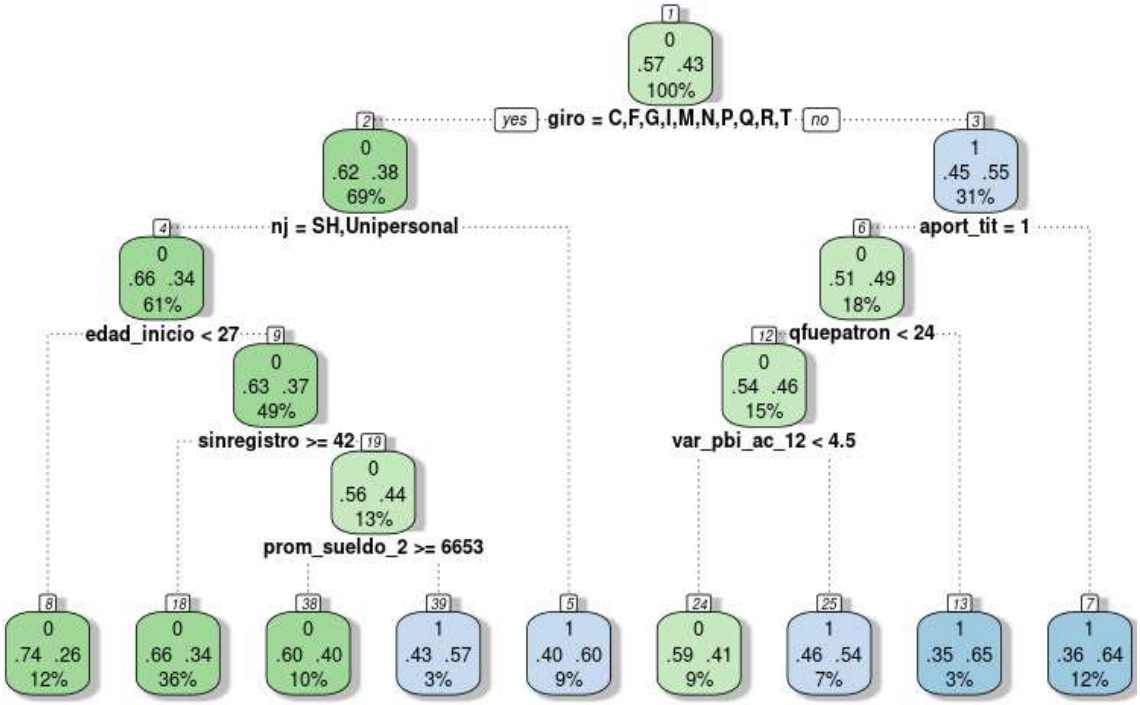
Recordando, inicialmente trabajamos con la Base con los 8513 registros que contienen la información de la historia laboral de los emprendedores y que es representativa de los 131.116 registros de emprendimientos iniciados entre 2010 y 2015.

4.4.1 – El árbol de clasificación

Al comenzar el análisis del árbol condicionamos el proceso a que el número mínimo de observaciones en los nodos intermedios sea el 9% de la Base, y en los terminales sea del 3%, de manera de obtener grupos mejor distribuidos.

Se exhibe el primer árbol que se conforma por seis niveles a partir de la base de datos de la muestra ya referenciada. El árbol representado por figuras de igual forma se encuentra diferenciado por los colores, donde la gama del azul destaca los grupos de emprendimientos que tienen más del 50% de probabilidad de lograr una sobrevida mayor a los 60 meses, y la de verde aquellos que tienen una probabilidad menor del 50% respectivamente.

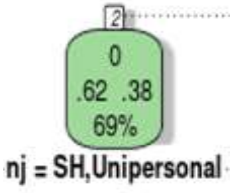
Figura 3 _ Árbol de clasificación de la base total de emprendimientos 2010-2015



Las variables que tienen mayor influencia para el estudio son, por orden de importancia: Giro, Naturaleza Jurídica y Tipo de aportación.

El primer nodo del árbol que conduce al nivel 1 expresa el 100% de los casos, donde explica que el 43% de los emprendimientos formales (iniciados en BPS) son los que sobreviven por un período de 5 y más años.

El Giro de la actividad del emprendimiento es una característica que impacta directamente en la sobrevivencia de estos, y es a partir de esta variable que se genera la primera división del árbol. Los emprendimientos que cumplen con la condición de la variable (especificada en negrita debajo de cada nodo) se agrupan a la izquierda del lector en la imagen (Sí cumplen), y los que no cumplen la condición se agrupan a la derecha.



Casi 7 de cada 10 emprendimientos provienen de los siguientes giros: **C** Industrias Manufactureras, **F** Construcción, **G** Comercio al por mayor y al por menor; reparación de los vehículos de motor y de las motocicletas, **I** Alojamiento y servicios de comida, **M** Actividades profesionales, científicas y técnicas, **N** Actividades administrativas y servicios de apoyo, **P** Enseñanza, **Q** Servicios sociales y relacionados con la Salud humana, **R** Artes, entretenimiento y recreación y **T** Actividades de los hogares en calidad de empleadores, actividades indiferenciadas de producción de bienes y servicios de los hogares. Este grupo de emprendimientos tiene una probabilidad de alcanzar 5 o más años de sobrevida en un 38%.

Con el mismo grupo de análisis, para realizar la siguiente distinción de los datos, la variable que se emplea es la Naturaleza Jurídica. El tipo de conformación legal del emprendimiento aparece como un diferenciador en cuanto a la sobrevida mostrada de la organización.
Casi 9 de cada 10 emprendimientos en este grupo cumplen la condición de ser Unipersonales o Sociedades de Hecho (en el total de los emprendimientos de la base representan el 61%). En cambio las Cooperativas,

SRL y otras formas jurídicas, presentan una mayor probabilidad de alcanzar 5 o más años de sobrevida, con un 60%. Como se ve del árbol, este grupo no presenta más divisiones (es el nodo "5" en el nivel 6 del árbol). En cambio, para el grupo mayoritario el algoritmo identifica la edad del emprendedor al inicio del emprendimiento como una variable influyente en el proceso, distinguiendo por un lado a un subgrupo cuyos emprendedores tienen menos de 27 años, obteniendo una probabilidad de éxito del 26%.
Para los casos donde los emprendedores tienen 27 o más años el algoritmo identifica la variable de "sin registro" como influyente, incluyendo a los emprendedores con más de 42 meses sin aportes a la seguridad social en un grupo con una probabilidad de éxito del 34%. Para el grupo de emprendedores que registran menos de 42 meses con huecos de nómina, el algoritmo vuelve a distinguir, esta vez por el promedio de sueldos del año anterior al inicio, diferenciando de acuerdo a un monto de $u 6.653[25]. Quienes percibieron sueldos mayores (en promedio) presentan una menor probabilidad de lograr el éxito con sus emprendimientos que aquellos que percibieron menos de dicho monto (40% y 57% respectivamente).

Volviendo al inicio del árbol, ahora centraremos el análisis en los emprendimientos que no cumplen la condición de la variable Giro. Es decir, analizaremos aquellos emprendimientos de los giros: **A** Producción agropecuaria, forestación y pesca; **H** Transporte y almacenamiento; **J** Información y comunicación; **K** Actividades financieras y de seguros; **L** Actividades inmobiliarias y **S** Otras actividades de servicio. Representan 3 de cada 10 emprendimientos, y en conjunto exhiben una mayor probabilidad de sobrevivida del 55% (respecto a los que cumplen la condición del Giro).

Continuando, en la siguiente distinción que realiza el algoritmo utiliza la variable Tipo de aportación, distinguiendo en un nodo terminal los emprendimientos vinculados al sector Rural, que representan el 12% de emprendimientos totales y tienen una probabilidad de sobrevida mayor a 5 años del 64%.

El otro grupo generado es con los de aportación Industria y Comercio, presentando un 49% de probabilidad de sobrevivir más de 5 años. En este caso, el algoritmo es capaz de incorporar otro nivel realizando la distinción de acuerdo a la variable 'Cantidad de meses que el emprendedor fue patrón' y el límite lo marca en 24 meses, denotando un pequeño subgrupo que supera dicho límite y representa el 3% del total de los emprendimientos, con un 65% de probabilidad de sobrevida a los 5 años.
Para los emprendedores que tienen una menor experiencia como patrones, el grupo conformado presenta un 46% de probabilidad de sobrevida a los 5 años, pero el algoritmo vuelve a distinguir, esta vez utilizando la variable de 'Variación del PBI acumulado último año' denotando una magnitud de 4.5 puntos. De acuerdo a esta distinción se generan 2 terminales, por un lado el sub grupo que en el período anterior a iniciar el emprendimiento la economía nacional había mostrado un mayor dinamismo, representando el 7% del total de emprendimientos con un 54% de probabilidad de sobrevida mayor a los 5 años. Por otro lado, los emprendimientos que iniciaron en un contexto de menor dinamismo económico, representando un 9% del total de emprendimientos y mostrando una probabilidad de 41% de sobrevida mayor a 5 años.

En síntesis, el 43% de los emprendimientos iniciados entre 2010 y 2015 presentaron una sobrevida de 5 o más años. Las características como el Giro, la Naturaleza jurídica y el tipo de aportación, así como las vinculadas a la historia laboral de los emprendedores (la experiencia como patrón), la edad del emprendedor al inicio, o incluso el desempeño de la economía en el período inmediato anterior, muestran una importante influencia que incide en la supervivencia de los emprendimientos.

[25] Valores 2014.

4.4.2 – La predicción del árbol de clasificación

La siguiente tabla presenta información relativa a los casos en que se producen los eventos (o no) de acuerdo a los 9 grupos (terminales) conformados por el árbol de clasificación.

El valor obtenido para IV es de 0.33, lo que indicaría que el predictor presenta una media-fuerte relación de los no-eventos/eventos. El coeficiente K-S presenta un valor de 25.84, lo que puede considerarse como aceptable para que el modelo sea explicativo.

Cuadro 2 _ Casos presentes y ausentes en el árbol de clasificación

Valor de información del modelo:			0,33				
Coeficiente K-S			25,84				
Nodos	Total	Ausentes	Presentes	Tot(%)	Aus(%)	Pres(%)	Prob_Pres(%)
13	330	115	215	3,88	2,46	5,61	65,15
7	1079	384	695	12,67	8,21	18,12	64,41
5	1168	453	715	13.72	9.68	18.64	61.22
39	267	120	147	3.14	2.57	3.83	55.06
25	546	250	296	6.41	5.34	7.72	54.21
24	722	393	329	8.48	8.40	8.58	45.57
38	775	471	304	9.10	10.07	7.93	39.23
18	2750	1835	915	32.30	39.23	23.86	33.27
8	876	657	219	10.29	14.04	5.71	25.00
Total	8513	4678	3835	100,00	100,00	100,00	no aplica

Nota: * los números de los Nodos son los correspondientes a los del último nivel de la Figura 4.
Para la confección de la tabla se emplea la función de predicción de "rpart" que devuelve información relativa a WOE, AP_ODD, IV y KS.

Como puede observarse de la predicción, el nodo 13 es el que presenta mayor probabilidad de casos presentes (favorables) de sobrevida de 5 o más años. Estamos refiriéndonos a los emprendimientos con aportación de Industria y Comercio, de los giros A, H, J, K, L y S, cuyos emprendedores tienen una experiencia como patrones de por lo menos 2 años (al momento de inicio del nuevo emprendimiento).

El otro grupo con similar probabilidad de casos favorables, corresponden a los mismos giros pero para los casos de la aportación Rural (el terminal 7 en el árbol).

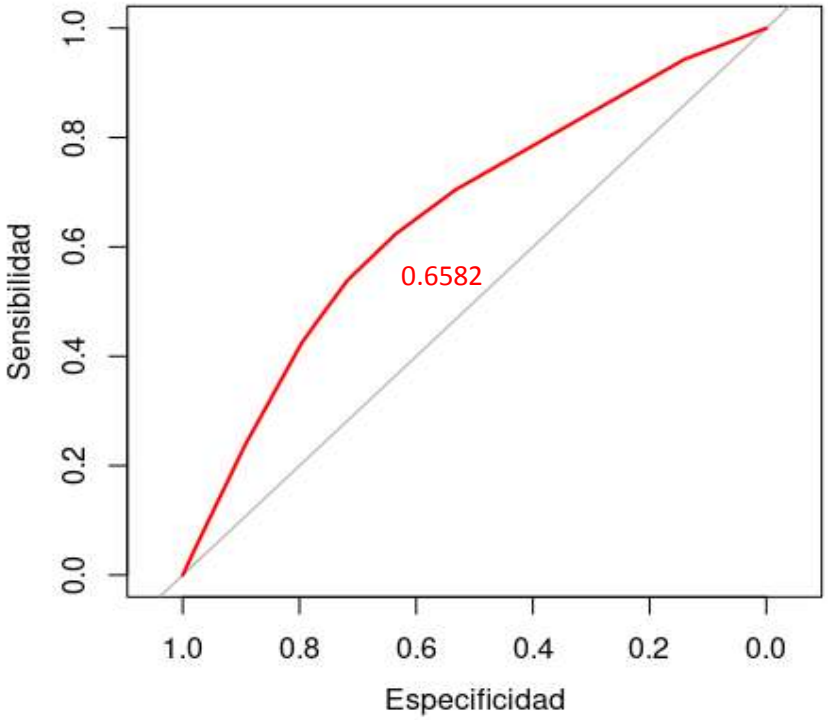
De acuerdo a la definición de WOE, a menor woe mayor es el peso del grupo para explicar el comportamiento de la variable dependiente. En nuestro ejemplo para el grupo con mayor probabilidad de casos presentes el cálculo del woe: "*ln*((115/4678)/(215/3835))" = -82,44. A mayor negatividad del woe, mejor explica el grupo la ocurrencia de la sobrevivencia.

En el Anexo C se presentan las salidas de los árboles para los casos de las bases filtradas según tengan o no dependientes, verificando que los casos que presentan trabajadores a cargo presentan una mayor sobrevida a los 5 años.

Luego de predecir la sobrevivencia a los 5 años, calculamos la curva Roc.

El valor alcanzado del área bajo la curva (AUC) es **0.6582**, por lo que puede considerarse como regular el test.

Gráfico 3 _ Curva Roc del árbol de clasificación



En definitiva, la clasificación realizada por el algoritmo y la predicción muestran una discreta performance del modelo.

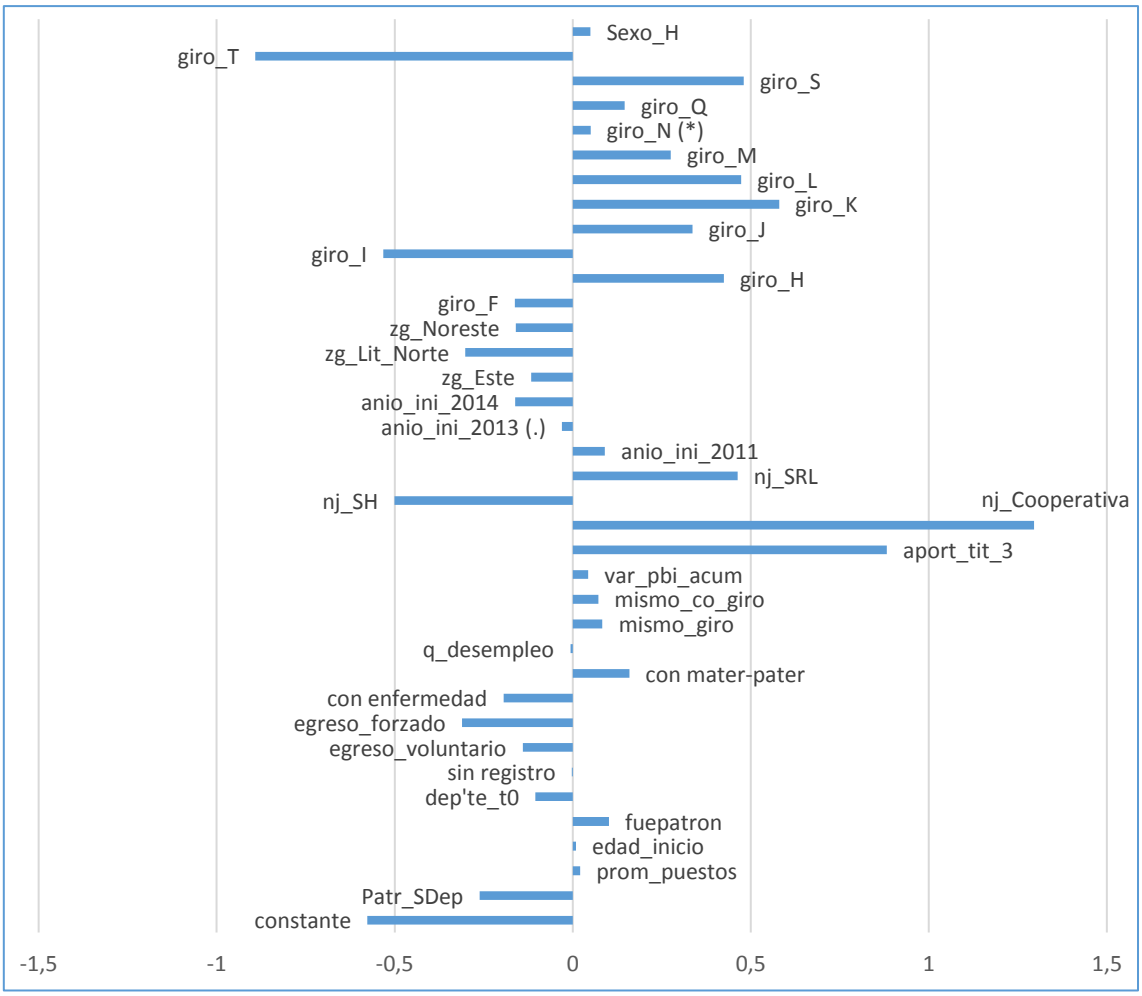
4.4.3 – La regresión logística

Empleando la misma base que para el proceso del árbol, pasamos ahora a realizar la regresión del modelo de sobrevivencia (en el Anexo D se comenta brevemente el proceso de la regresión, así como las variables del modelo original).

Recordemos que lo que se estima en la regresión logística no es Y', sino un Logit, es decir, el logaritmo de la probabilidad de que ocurra un evento, frente a la probabilidad de que no ocurra.

El siguiente gráfico ayuda a obtener una mejor interpretación de los resultados obtenidos en la salida logit.

Gráfico 4 _ Representación del signo de los coeficientes en la regresión logística.



Niveles de significación en el gráfico: sin marca : [0 ; 0,001], (*) : (0,01 ; 0,05], (.) : (0,05 ; 0,1]

Como puede notarse, casi todas las variables son significativas al 99% (sólo dos tienen una significancia algo menor).

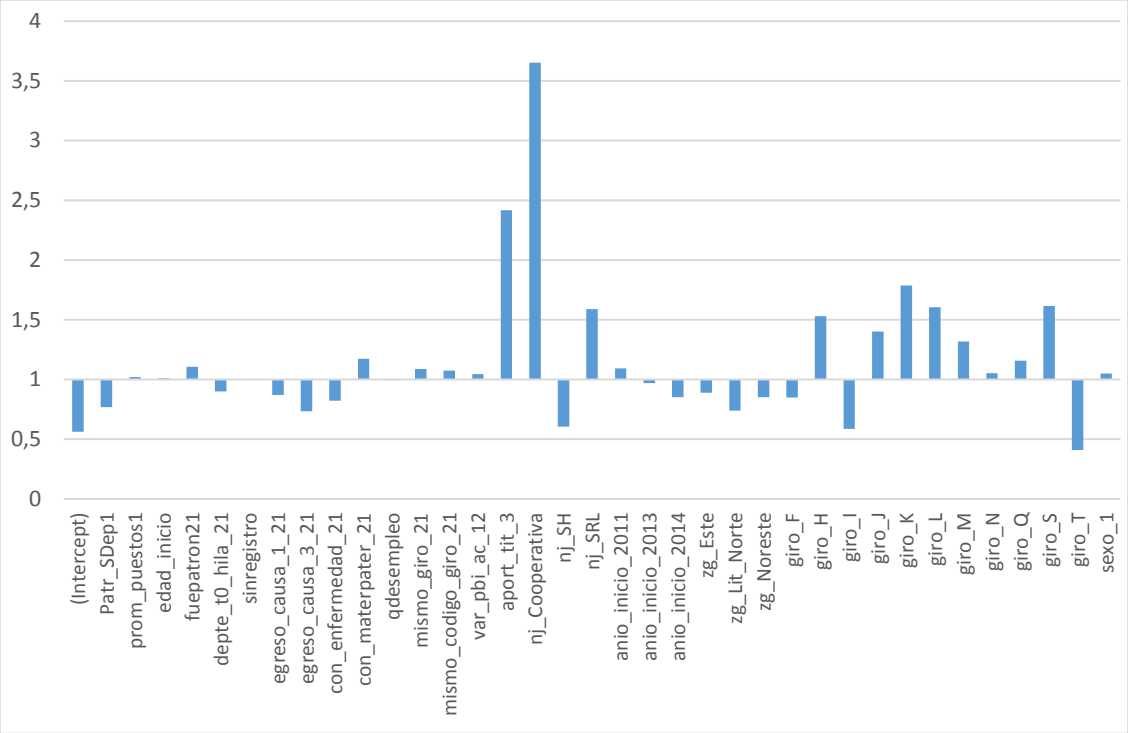
Las barras que en el gráfico se expanden hacia la derecha del lector (valores positivos) se representan de acuerdo al signo positivo en la salida de la regresión, y cuanto mayor sea su valor, presentar dicha característica significará una ventaja frente a la probabilidad de ocurrencia de que el emprendimiento Sobreviva a los 5 años.

Las que se representan en sentido opuesto (valores negativos) son las de signo negativo en la salida de la regresión, y cuanto menor sea su valor presentar dicha característica significará una desventaja frente a la probabilidad de ocurrencia de que el emprendimiento Sobreviva a los 5 años.

La comparación siempre debe hacerse en función de que, manteniendo el resto de las variables constantes, la variable de análisis influye positiva o negativamente de acuerdo al signo que presente.

A su vez, la interpretación de los coeficientes debe hacerse de acuerdo al Odd-ratio, es decir a “exp(β_i)”, cuanto más diferente a 1 sea este valor implicará que cualquier incremento en los niveles de la variable “i” tendrá un efecto significativo sobre la variable dependiente.

Gráfico 5 _ Odds-ratios de las variables en la regresión logística.



Puede observarse que la variable que presenta la mayor influencia positiva en la probabilidad de Sobrevivencia de los emprendimientos es la "nj_Cooperativa". El valor del odd ratio para dicha variable es 3,65 aproximadamente, la interpretación que puede hacerse es que para las Cooperativas el ratio de las probabilidades de sobrevivir los 5 años aumenta entre 3 y 4 veces.

Otras variables con notoria influencia positiva son "tipo de aportación 3", "giro_K", "giro_S", "giro_L", "nj_SRL" , "giro_H", "giro_J" y "giro_M", con valores de odd-ratio 2,42 – 1,79 – 1,62 – 1,60 – 1,59 – 1,53 – 1,40 y 1,32 respectiva y aproximadamente[26].

Demás características que influyen positivamente pero con menor relevancia que las anteriores son "con maternidad-paternidad", "fuepatrón", "giro_Q", "mismo_giro", "mismo_código_giro" y "anio_inicio_2011".

En el sentido opuesto, la variable que presenta la mayor influencia negativa en la probabilidad de Sobrevivencia es el "giro_T" (Actividades de los hogares en calidad de empleadores, etc), con un odd-ratio de 0,41 aproximadamente, la interpretación que puede hacerse es que para estas Actividades el ratio de las probabilidades de sobrevivir los 5 años disminuye entre 2 y 3 veces.

Otras variables con notoria influencia negativa son "giro_I", "nj_SH", "egreso_causa_3", "zg_Lit_Norte" y "Patr_Sdep", con valores de odd-ratio 0,59 – 0,61 – 0,73 – 0,74 y 0,77 respectiva y aproximadamente[27].

[26] "tipo de aportación 3" - Rural, "giro_K" - Actividades financieras y de seguros, "giro_S" - otras actividades de servicio, "giro_L" - Actividades inmobiliarias, "nj_SRL" – Sociedad de responsabilidad limitada, "giro_H" - Transporte y almacenamiento, "giro_J" - Información y comunicación y "giro_M" - Actividades profesionales, científicas y técnicas.
[27] "giro_I" - Alojamiento y servicios de comida, "nj_SH" – Sociedad de Hecho, "egreso_causa_3" – egreso forzado (despido, término contrato), "zg_Lit_Norte" – Artigas, Salto, Paysandú y "Patr_Sdep" – patrón sin dependientes.

DEMÁS características que influyen negativamente pero con menor relevancia que las anteriores son “con enfermedad”, “egreso causa 1”, “anio_inicio_2014”, “zg_Noreste” y “giro_F”.

4.4.4 – La predicción de la regresión logística

De manera análoga al análisis realizado en la predicción del árbol, en el siguiente cuadro se presenta información relativa a los casos en que se producen los eventos (o no) de acuerdo a los deciles conformados por la regresión.

El valor obtenido para IV es de 0.39, lo que indicaría que el predictor presenta una media-fuerte relación de los no-eventos/eventos. El coeficiente K-S presenta un valor de 25.85, lo que puede considerarse como aceptable para que el modelo sea explicativo.

Cuadro 3 _ Casos presentes y ausentes en la regresión.

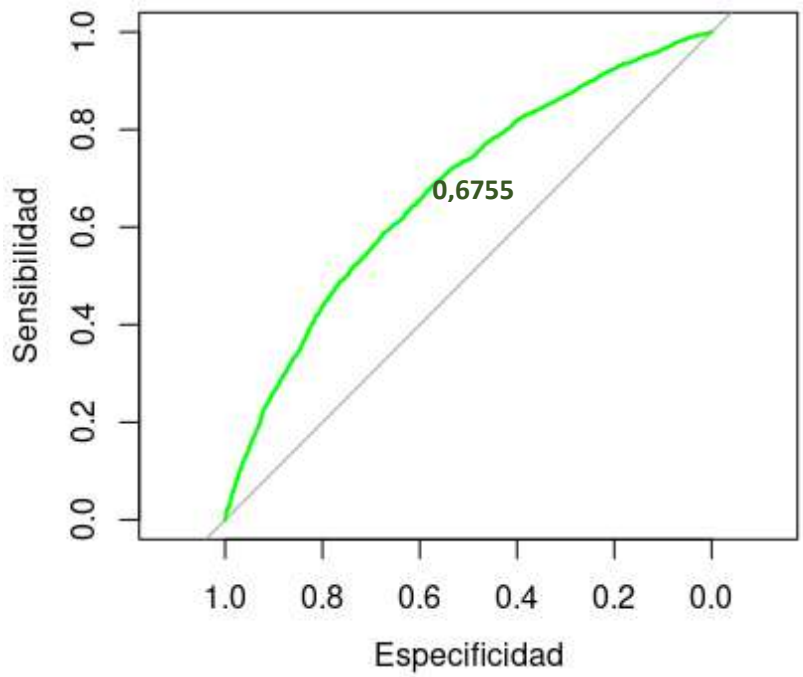
Valor de información del modelo:			0,39				
Coeficiente K-S			25,85				
Decil *	Total	Ausentes	Presentes	Tot(%)	Aus(%)	Pres(%)	Prob_Pres(%)
10	851	249	602	10,00	5,32	15,70	70,74
9	851	315	536	10,00	6,73	13,98	62,98
8	851	348	503	10,00	7,44	13,12	59,11
7	852	424	428	10,01	9,06	11,16	50,23
6	851	458	393	10,00	9,79	10,25	46,18
5	851	496	355	10,00	10,60	9,26	41,72
4	852	523	329	10,01	11,18	8,58	38,62
3	851	596	255	10,00	12,74	6,65	29,96
2	851	610	241	10,00	13,04	6,28	28,32
1	852	659	193	10,01	14,09	5,03	22,65
Total	8513	4678	3835	100,00	100,00	100,00	no aplica

Nota: * los números expresan la apertura por deciles de la Base.
Para la confección de la tabla se emplea la función de predicción de “rpart” que devuelve información relativa a WOE, AP_ODD, IV y KS.

El último decil presenta la mayor probabilidad de casos presentes (favorables), aproximadamente 7 de cada 10. La mayoría de los deciles (6) tienen menos de la mitad de casos favorables.
En la regresión el grupo con mayor probabilidad de casos presentes el cálculo del woe: “*ln*((249/4678)/(602/3835))” = -108,15. A mayor negatividad del woe, mejor explica el grupo la ocurrencia de la sobrevivencia.

Luego de predecir la sobrevivencia a los 5 años, calculamos la curva Roc.
El valor alcanzado del Área bajo la curva es **0.6755**, por lo que también en este caso puede considerarse como regular el test.

Gráfico 6 _ Curva Roc de la regresión logística



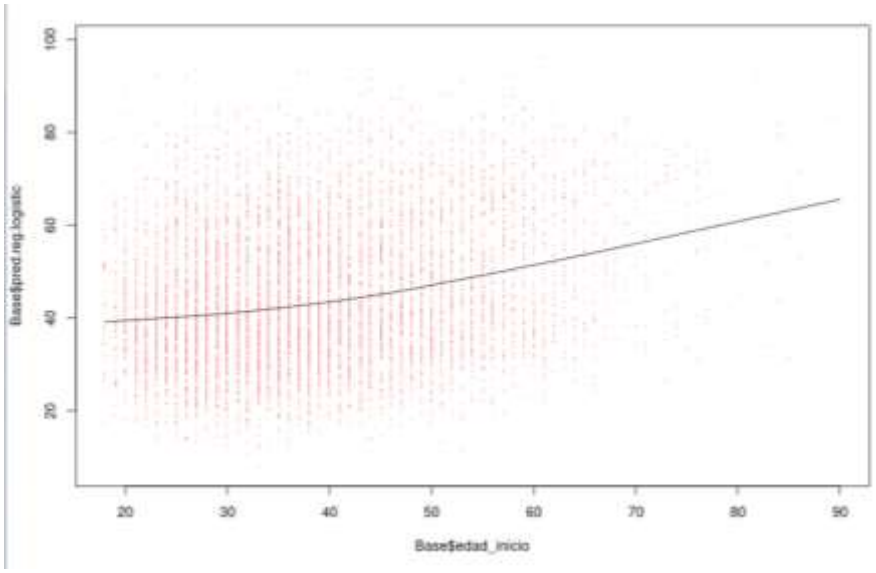
También en este caso la performance del modelo es discreta. En el Anexo F se exhibe de manera sucinta los pasos de todo el proceso.

4.4.5 – Supervivencia en función de variables clave

Realizamos algunos cruces para obtener una mejor comprensión del comportamiento que presentan ciertas variables. Para facilitar el análisis visual, en el eje vertical siempre se expresa la probabilidad de éxito del emprendimiento (60 meses o más de supervivencia).

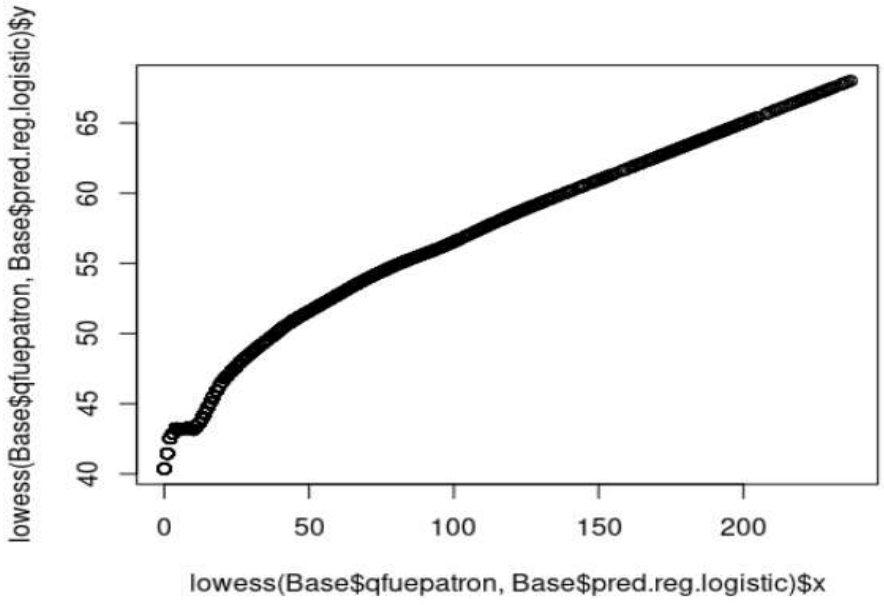
Si bien el análisis se realizó sin distinguir entre seniors y no seniors, es interesante ver lo que sucede con la variable de Edad. Observemos como varía la probabilidad de que un emprendimiento sobreviva 5 o más años, de acuerdo a la edad que presente el emprendedor al inicio del mismo. Sumar más años de vida tiene una influencia positiva, pero además puede observarse que a partir de los 40 - 50 años esta relación se hace más notoria. Si bien en la salida de la regresión se obtiene un coeficiente pequeño para esta variable, lo que significaría una pequeña influencia, la misma refiere a la diferencia de un año de edad en el emprendedor. En el gráfico lo que visualizamos sería el efecto acumulado arrojado por la variable.

Gráfico 7 _ Sobrevivencia según la edad del emprendedor al inicio.



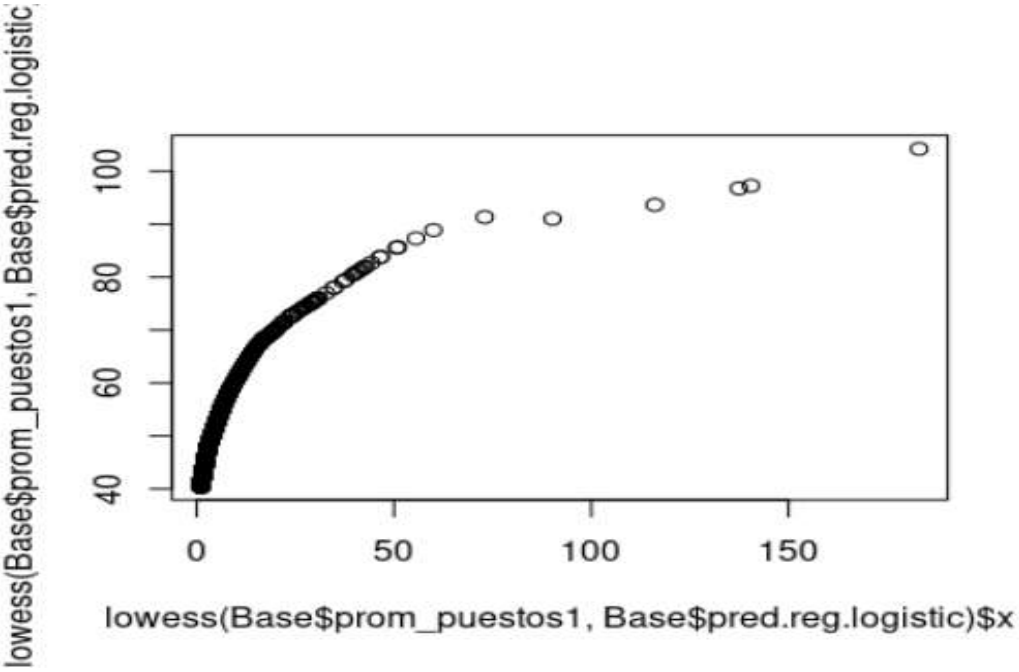
Como vimos también, a medida que aumenta la experiencia de los emprendedores como patrones, crece la probabilidad de que sus emprendimientos logren el éxito de sobrevivir. En un primer período de aproximadamente 1 año (12 meses) la influencia no aparece clara, pero posteriormente la relación queda explícita.

Gráfico 8 _ Sobrevivencia según la experiencia como patrón del emprendedor.



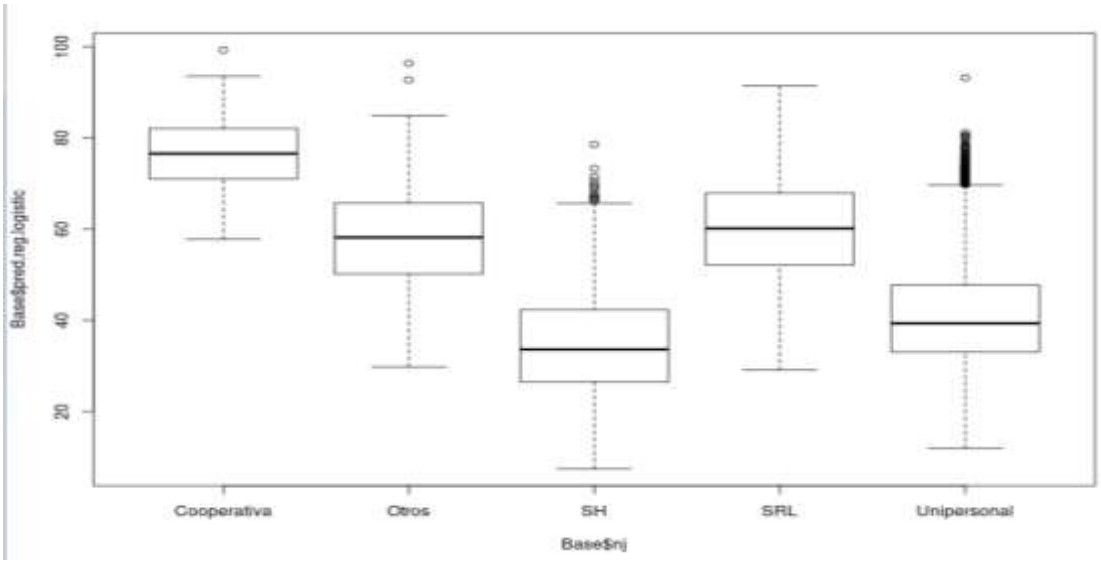
La condición de tener o no trabajadores a cargo en el emprendimiento influye en la variable dependiente. En el entendido de que un emprendimiento con muchos trabajadores dependientes sugiere una planificación, inversión y visión de mediano o largo plazo, y de que pueda existir entonces un mayor respaldo ante situaciones adversas del ciclo económico, es coherente esperar este comportamiento respecto al tamaño de la empresa. Si bien lo anterior no excluye la posibilidad de que los emprendimientos sin dependientes puedan ser exitosos, la cantidad de puestos es una variable a tener en cuenta.

Gráfico 9 _ Sobrevivencia según la cantidad de puestos dependientes del emprendimiento.



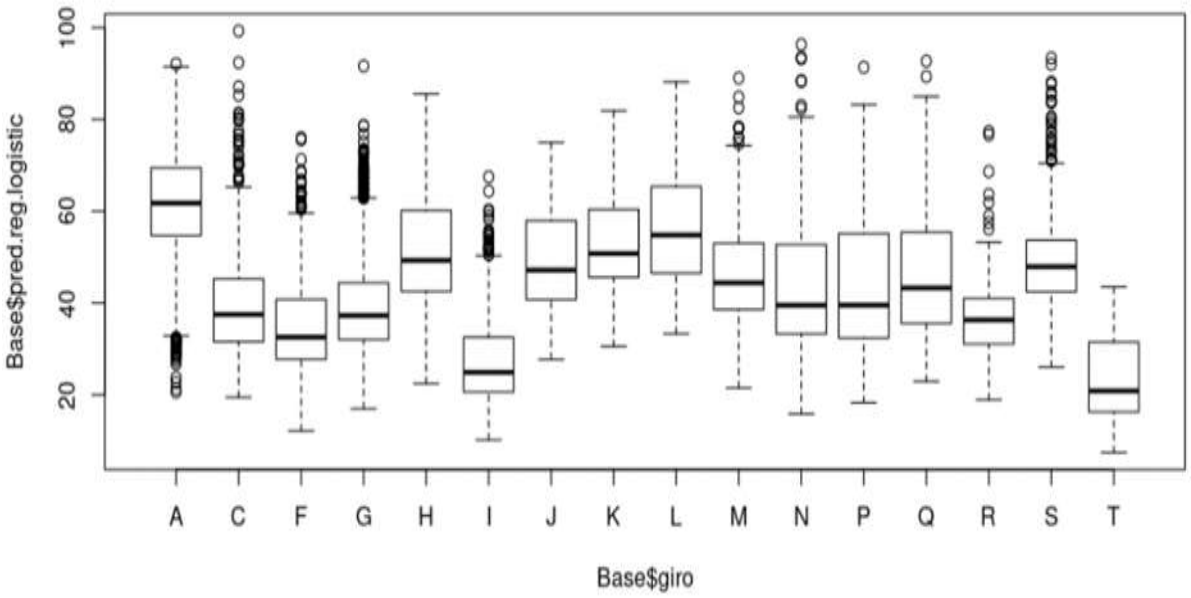
Generamos otro cruce con la naturaleza jurídica del emprendimiento. Del gráfico de caja se observa que en promedio los casos de Sociedades de hecho o Unipersonales presentan una menor probabilidad de alcanzar una sobrevida mayor a los 5 años. En el entendido que, en general, se trata de emprendimientos sin dependientes, y por lo que constatamos de los resultados la cantidad de puestos influye positivamente en la sobrevida, es lógico que en promedio estos tipos de NJ tengan más dificultades para perdurar en el tiempo.

Gráfico 10 _ Sobrevivencia según la Naturaleza Jurídica del emprendimiento



El último cruce presentado es acerca del tipo de actividad (Giro) del emprendimiento. Las actividades que influyen de manera diferente al resto son, positivamente A, L, K, y negativamente T, I y F.

Gráfico 11 _ Sobrevivencia según la actividad económica del emprendimiento



A Producción agropecuaria, forestación y pesca; **K** Actividades financieras y de seguros; **L** Actividades inmobiliarias.
F Construcción, **I** Alojamiento y servicios de comida, **T** Actividades de los hogares en calidad de empleadores, actividades indiferenciadas de producción de bienes y servicios de los hogares.

5. Síntesis

El Emprendedurismo está asociado a una forma de crear bienes y servicios en la cultura de la cooperación e innovación que desafían al mercado incrementando la productividad a partir de identificar las oportunidades de negocio que se generan en el entorno. Como ejemplo de actividades que emprenden podemos mencionar entre otras: Especialistas en el desarrollo de ideas en comunicación y marketing, Pastelería artesanal, Venta de plantas, Soporte y servicio técnico informático integral para la tercera edad, Diseño de productos realizados con madera y materiales reciclados/reutilizados, Venta de chocolate, Confección de prendas a medida y Cursos online de costura.

A partir de una muestra de los emprendimientos formales registrados en BPS de las aportaciones Rural e Industria y Comercio en el período 2010 - 2015, el objetivo de esta investigación fue identificar las variables que presentan una influencia significativa para explicar la mayor sobrevivencia de éstos.

El criterio utilizado para la medición de la supervivencia se basó en investigaciones internacionales y nacionales donde se evidencia que los primeros años de vida son críticos para los emprendimientos, logrando una mayor estabilidad a partir de los 5 años (60 meses y más).

Fueron dos técnicas las que se emplearon a través del uso del software R para realizar el análisis, primero por medio de árboles de clasificación y luego por medio de un modelo de regresión logística.

Dentro de las principales variables, ambas técnicas identifican varios giros de actividad, algunos tipos de naturalezas jurídicas y al tipo de aportación como las de mayor influencia para explicar las diferencias en la sobrevivencia entre unos emprendimientos y otros.

En lo que respecta a las características de los emprendimientos surge de esta investigación que aquellas actividades vinculadas a la producción agropecuaria, financieras y de seguros, tecnológicas y científicas presentan una mayor probabilidad de incrementar la sobrevivencia. Este resultado reviste coherencia, si se tiene en cuenta el caso de las actividades agropecuarias, éstas se constituyen en general en proyectos de largo plazo; en tanto para los giros relacionados con la tecnología y ciencia, éstos se inscriben dentro de las actividades de mayor impulso reciente.

En la misma línea los emprendimientos que se constituyen como cooperativas también conllevan una mayor sobrevida. Esto podría estar implícito en el entendido que están relacionadas a cubrir necesidades de vivienda y consumo, presentado además una forma de asociación particular. Por otro lado, los emprendimientos que al inicio cuentan con al menos un dependiente también muestran una mayor sobrevivencia, lo que podría explicar que la incorporación de puestos de trabajo se asocia a una inversión previa más importante dentro de una idea de negocio a mediano o largo plazo.

Con relación a las características del emprendedor también se realizan distinciones de acuerdo a la edad, a la experiencia como patrón, a la discontinuidad en la aportación a la seguridad social de éstos, o a la causa del egreso en el trabajo anterior del emprendedor. Estas son variables que influyen significativamente en la sobrevida de los emprendimientos.

Surge de los resultados que los emprendedores de mayor edad consiguen extender sus emprendimientos por más tiempo, y en el mismo sentido, la experiencia previa como patrón en una organización, o en el sector de actividad influyen en esta misma dirección. Es esperable que el conocimiento del negocio, además de la acumulación de vivencias, colaboren a extender la duración de la organización.

De forma complementaria se tuvo en consideración el contexto económico, marcando una influencia positiva la variación acumulada del PBI en el año anterior al inicio del emprendimiento. Se desprende entonces que un ciclo económico favorable influye en la decisión de emprender, así como en lograr su estabilización.
Por último, también se encuentran diferencias aunque leves respecto a la sobrevivencia del emprendimiento, de acuerdo a la ubicación geográfica declarada. Se observa en promedio una mayor probabilidad de sobrevivencia en el litoral sur (que comprende a los departamentos de Colonia, Rio Negro y Soriano), donde podría estar influyendo, además de la productividad de las tierras que atrae a importantes desarrollos agroindustriales, la fortaleza de las instituciones acompañadas de la cultura emprendedora local.

Para los mejores modelos que construimos empleando cada una de las técnicas referidas, tanto los indicadores en las tablas de pronósticos de sobrevivencia como los valores de las áreas bajo la curva Roc, denotan que el comportamiento de la variable dependiente (Supervivencia_T5) en función del resto se explica sólo de manera aceptable.

En vista de ello, es menester recordar que para el análisis realizado no se contó con información específica relativa a la gestión de los emprendimientos, tales como la facturación mensual, la inversión realizada para

iniciar el proyecto (además sin el acceso a otros datos como el origen de los fondos, la solicitud de un crédito o el apoyo de redes familiares y/o sociales para emprender), que son relevantes para analizar la incidencia en la duración de cualquier proyecto.

En resumen ambas técnicas devuelven resultados similares y explican de manera satisfactoria (teniendo en cuenta las restricciones de información) la manera en que influyen las principales variables en la evolución de la supervivencia, lo que fortalece las conclusiones obtenidas.

En este estudio la variable dependiente fue la supervivencia a 5 años; en un futuro se pueden plantear otros horizontes temporales de supervivencia, o cambiar el foco de interés de la investigación, de forma de enriquecer y complementar el análisis.

6. – Bibliografía

"Emprendedurismo Senior en Uruguay: el envejecimiento como una nueva oportunidad para el crecimiento", Antúnez, Naranja y Nuñez, AGSS-BPS, Comentarios de Seguridad Social No. 74, abril 2021.

"Emprendedurismo Senior en Uruguay. Caracterización y análisis de los emprendedores afiliados a BPS, 2010-2019", Antúnez, Naranja y Nuñez, AGSS-BPS, Comentarios de Seguridad Social No. 82, diciembre 2021.

"Aplicación de árboles de decisión para la estimación del escenario económico y la estimación de movimiento la tasa de interés en Chile", Carlos Dupouy Berrios, Tesis para optar al grado de Magíster en Finanzas. Profesor guía David Díaz Solís Ph.D., Santiago de Chile, julio 2014.

"Econometría de datos de panel en R: el paquete plm", Yves Croissant, Giovanni Millo, Diario de Sta Software estadístico, Universidad de California, Los Angeles, 2008.

"Regresión Logística", Universidad ORT Uruguay, 2020. Presentación elaborada tomando como base el capítulo 4 de G. James et. al., An Introduction to Statistical Learning: with Applications in R. New York :Springer, 2013.

"Un paquete R para análisis masivos de modelos predictivos de regresión logística multivariante, y sus medidas de discriminación y de clasificación asociadas", Balsa Carlos, Sanchez Alexandre, UOC, 2017.

Links consultados:

https://rpubs.com/jboscomendoza/arboles_decision_clasificacion
https://analisisdedatos.net/mineria/tecnicas/arbolesDecision/rpart.php
https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/
https://rpubs.com/Joaquin_AR/229736
https://rpubs.com/elfenixsoy/arbol-veronica#:~:text=Los%20%C3%A1rboles%20de%20decisi%C3%B3n%20es,problemas%20(Beltr%C3%A1n%2C%202015)
https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=dm-crisp-help-overview

https://www.sngular.com/es/data-science-crisp-dm-metodologia/
https://www.cienciadedatos.net/documentos/27 regresion logistica simple y multiple
http://www.hrc.es/bioest/Reglog 1.html
https://www.sciencedirect.com/science/article/pii/S021391110371694X
https://arcruz0.github.io/libroadp/logit.html

7. Anexo

A _ Variables incluidas en la Base_

"id_empr_tit" _ identificador de emprendimiento
"Patr_SDep" _ dummy, 1 si el emprendimiento no tiene dependientes
"anio_inicio" _ año de inicio del emprendimiento
"nj" _ naturaleza jurídica del emprendimiento
"zg" _ zona geográfica donde se localiza el emprendimiento
"Y_36" _ dummy, 1 si el emprendimiento tiene al menos 36 meses de sobrevida
"Y_42" _ dummy, 1 si el emprendimiento tiene al menos 42 meses de sobrevida (empresa) "Y_60" _ dummy, 1 si el emprendimiento tiene al menos 60 meses de sobrevida "prom_masasal1" _ masa salarial promedio que paga el emprendimiento en el primer año
"prom_puestos1" _ promedio de puestos que tiene el emprendimiento en el primer año
"mes_inicio" _ mes del año en que inicia el emprendimiento
"trunc" _ dummy, 1 si al final del período sigue activo el emprendimiento
"mueren_T1" _ dummy, 1 si el emprendimiento muere dentro del primer año
"mueren_T2" _ dummy, 1 si el emprendimiento muere dentro del segundo año
"mueren_T3" _ dummy, 1 si el emprendimiento muere dentro del tercer año
"mueren_T4" _ dummy, 1 si el emprendimiento muere dentro del cuarto año
"mueren_T5" _ dummy, 1 si el emprendimiento muere dentro del quinto año
"sobreviven_T1" _ dummy, 1 si el emprendimiento sobrevive al primer año
"sobreviven_T2" _ dummy, 1 si el emprendimiento sobrevive al segundo año
"sobreviven_T3" _ dummy, 1 si el emprendimiento sobrevive al tercer año
"sobreviven_T4" _ dummy, 1 si el emprendimiento sobrevive al cuarto año
"sobreviven_T5" _ dummy, 1 si el emprendimiento sobrevive al quinto año
"ERC" _ dummy, 1 si el emprendimiento es Empresa de Rápido Crecimiento
"ERC_OCDE" _ dummy, 1 si el emprendimiento es ERC según criterios de OCDE
"pers_identificador" _ identificador del emprendedor
"t0" _ momento de inicio del emprendimiento
"sexo" _ sexo del emprendedor
"edad_inicio" _ edad del emprendedor al inicio del emprendimiento
"fuepatron2" _ dummy, 1 si el emprendedor fue patrón en algún momento de su vida pasada
"qfuepatron" _ cantidad de meses que en el pasado el emprendedor fue patrón
"depte_t0_hila_2" _ dummy, 1 si el emprendedor iniciar también es dependiente en otra empresa
"no_depte_t0_hila_2" _ dummy, 1 si el emprendedor al iniciar también es no dep'te en otra empresa
"sinregistro" _ cantidad de meses sin aportes a la seguridad social en la historia de los emprendedores
"egreso_causa_1_2" _ dummy, 1 si en el año previo al inicio el emprendedor tiene baja voluntaria
"egreso_causa_2_2" _ dummy, 1 si en el año previo al inicio el emprendedor tiene baja por jubilación
"egreso_causa_3_2" _ dummy, 1 si en el año previo tiene baja por despido o término de contrato
"con_desempleo_2" _ dummy, 1 si en 2 años previos el emprendedor presenta seguro por desempleo
"con_enfermedad_2" _ dummy, 1 si en 2 años previos el emprendedor presenta seguro por enfermedad
"con_materpater_2" _ dummy, 1 si en 2 años previos el emprendedor usa benef. de maternidad o pater.

"qdesempleo" _ cantidad de meses que el emprendedor presentó seguro de desempleo previo al inicio "prom_sueldo_2" _ promedio de sueldos que tiene el emprendedor en el año previo al inicio "mismo_giro_2" _ dummy, 1 si el emprendedor tiene experiencia en la actividad que emprende "mismo_codigo_giro_2" _ dummy, 1 si el emprendedor tiene experiencia en el tipo de emprendimiento
"muestra" _ dummy, 1 si tiene información de Historia Laboral
"exp" _ expansor, ponderador de la muestra al total de la base
"tasa_desempleo" _ valor de la TD por mes
"tasa_empleo" _ valor de la TE por mes
"var_pbi_ac_12" _ valor de la variación del PIB acumulado anual
"aport_tit" _ tipo de aportación del emprendimiento
"codigo_de_giro" _ CIIU del emprendimiento
"giro" _ tipo de actividad del emprendimiento
"aport_tit_3" _ dummy, 1 si es un emprendimiento Rural
"nj_Cooperativa" _ dummy, 1 si el emprendimiento es una Cooperativa
"nj_SH" _ dummy, 1 si el emprendimiento es una Sociedad de Hecho
"nj_SRL" _ dummy, 1 si el emprendimiento es una Sociedad de Responsabilidad Limitada
"nj_Unipersonal" _ dummy, 1 si el emprendimiento es una Unipersonal
"anio_inicio_2011" _ dummy, 1 si el emprendimiento inicia en 2011
"anio_inicio_2012" _ dummy, 1 si el emprendimiento inicia en 2012
"anio_inicio_2013" _ dummy, 1 si el emprendimiento inicia en 2013
"anio_inicio_2014" _ dummy, 1 si el emprendimiento inicia en 2014
"anio_inicio_2015" _ dummy, 1 si el emprendimiento inicia en 2015
"mes_inicio_1" _ dummy, 1 si el emprendimiento inicia en enero
"mes_inicio_2" _ dummy, 1 si el emprendimiento inicia en febrero
"mes_inicio_3" _ dummy, 1 si el emprendimiento inicia en marzo
"mes_inicio_4" _ dummy, 1 si el emprendimiento inicia en abril
"mes_inicio_5" _ dummy, 1 si el emprendimiento inicia en mayo
“mes_inicio_6" _ dummy, 1 si el emprendimiento inicia en junio
"mes_inicio_8" _ dummy, 1 si el emprendimiento inicia en agosto
"mes_inicio_9" _ dummy, 1 si el emprendimiento inicia en setiembre
"mes_inicio_10" _ dummy, 1 si el emprendimiento inicia en octubre
"mes_inicio_11" _ dummy, 1 si el emprendimiento inicia en noviembre
“mes_inicio_12" _ dummy, 1 si el emprendimiento inicia en diciembre
"zg_Este" _ dummy, 1 si el emprendim. se ubica en el este (Treinta y Tres, Lavalleja, Rocha, Maldonado)
"zg_Lit_Norte" _ dummy, 1 si el emprendimiento se ubica en el litoral norte (Artigas, Salto, Paysandú)
"zg_Lit_Sur" _ dummy, 1 si el emprendimiento se ubica en el litoral sur (Río Negro, Soriano, Colonia)
“zg_Metropolitana" _ dummy, 1 si se ubica en la zona metropolitana (Canelones, Montevideo, San José)
"zg_Noreste" _ dummy, 1 si el emprendim. se ubica en el noreste (Cerro Largo, Tacuarembó, Rivera)
"giro_A" _ dummy, 1 si el giro es ‘Producción agropecuaria, forestación y pesca’
"giro_F" _ dummy, 1 si el giro es ‘Construcción’
"giro_G" _ dummy, 1 si el giro es ‘Comercio al por mayor y al por menor; reparación de los vehículos de motor y de las motocicletas’
"giro_H" _ dummy, 1 si el giro es ‘Transporte y almacenamiento’
"giro_I" _ dummy, 1 si el giro es ‘Alojamiento y servicios de comida’
"giro_J" _ dummy, 1 si el giro es ‘Información y comunicación’
"giro_K" _ dummy, 1 si el giro es ‘Actividades financieras y de seguros’
"giro_L" _ dummy, 1 si el giro es ‘Actividades inmobiliarias’
"giro_M" _ dummy, 1 si el giro es ‘Actividades profesionales, científicas y técnicas’
"giro_N" _ dummy, 1 si el giro es ‘Actividades administrativas y servicios de apoyo’

"giro_P" _ dummy, 1 si el giro es 'Enseñanza'
"giro_Q" _ dummy, 1 si el giro es 'Servicios sociales y relacionados con la salud humana'
"giro_R" _ dummy, 1 si el giro es 'Arte, entretenimiento y educación'
"giro_S" _ dummy, 1 si el giro es 'Otras actividades de servicio'
"giro_T" _ dummy, 1 si el giro es 'Actividades de los hogares en calidad de empleadores, actividades indiferenciadas de producción de bienes y servicios de los hogares para uso propio'
"sexo_1" _ dummy, 1 si el emprendedor es hombre

Nota: las dummys de la historia laboral refieren al momento del inicio del nuevo emprendimiento

B _ Árbol de decisión_

La implementación particular de CART que usaremos es conocida como **Recursive Partitioning and Regression Trees** o **RPART**. De allí el nombre del paquete que utilizaremos en nuestro análisis.
La función "rpart" generará un árbol proporcionándole una fórmula, indicándole los datos que debe usar y estableciendo el método (de clasificación en nuestro caso). Este es un algoritmo no supervisado que puede encontrar grupos ocultos en los datos, o intuidos pero no etiquetados. Una característica muy importante en este algoritmo es que una vez que alguna variable ha sido elegida para separar los datos, ya no es usada de nuevo en los grupos que ha creado. Se buscan variables distintas que mejoren la separación de los datos.
La función "rpart" que se utiliza hace crecer el árbol deteniéndose cuando cierto criterio se alcanza. Dentro de la función pueden especificarse los parámetros "Minsplit" y "minbucket", estos son –respectivamente- los números mínimos de observaciones en un nodo intermedio (para particionarlo) y en un nodo terminal.

El árbol para de crecer cuando:

1_ el decremento de la desviación va por debajo de cierto umbral
2_ el número de muestras en el nodo es menor que otro umbral
3_ la profundidad del árbol excede otro valor

(Los umbrales son controlados por los parámetros cp (1), minlist (2) y maxdepth (3). Por defecto estos valores son 0.01, 20 y 30, respectivamente. Si se desea evitar el problema del sobreajuste se debe verificar la validez de estos criterios.)

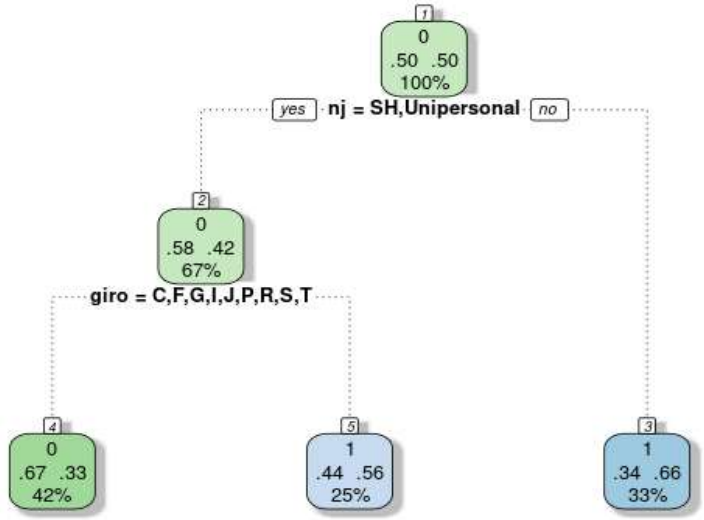
El paquete rpart implementa un método para podar llamado costo de complejidad de podar. Este método utiliza el valor del parámetro "cp" que R calcula para cada nodo del árbol. Este método para podar trata de estimar el valor de cp que asegura el mejor compromiso entre la precisión predictiva y el tamaño del árbol.
Dado un árbol obtenido con la función rpart(), R puede producir un conjunto de sub-árboles de este árbol y estimar su desempeño predictivo. Esta información puede ser obtenida utilizando la función printcp().
Una vez hecho esto, los datos son separados (particionados) en grupos a partir de la regla obtenida. Después, para cada uno de los grupos resultantes, se repite el mismo proceso. Se busca la variable que mejor separa los datos en grupos, se obtiene una regla, y se separan los datos. Hacemos esto de manera recursiva hasta que nos es imposible obtener una mejor separación. Cuando esto ocurre, el algoritmo se detiene. Cuando un grupo no puede ser partido de una mejor manera, se le llama nodo terminal u hoja.

C _ Generación del árbol distinguiendo la presencia de trabajadores a cargo

Base filtrada para los emprendimientos con dependientes

Para el proceso, se realizaron todos los ajustes necesarios para no generar inconvenientes a la hora de procesar en el software R. Según los requerimientos previstos para modelar, fue necesario modificar el formato de algunas variables, para poder obtener un mejor resultado del proceso del árbol de decisión.
Al realizar el análisis del árbol para la base de emprendimientos únicamente para los que presentan trabajadores dependientes, surge que las variables que el algoritmo selecciona para distinguir los diferentes grupos, cambian su relevancia. La principal característica para distinguir grupos ahora pasa a ser la **Naturaleza Jurídica**, y luego la sigue el **Giro** (recordemos que el Giro era la principal cuando analizamos la Base completa). En este caso, el árbol no realiza más particiones.

Gráfico 12 _ Árbol de clasificación de la base con dependientes



Este árbol muestra en el primer nodo que el total de emprendimientos (con dependientes) presenta una sobrevida de 5 o más años del 50%.

Como puede observarse, un tercio aproximadamente de estos emprendimientos no cumplen la condición de ser sociedades de hecho o unipersonales, y son los que presentan la mayor probabilidad de éxito dentro de los 3 nodos terminales (2 de cada 3 emprendimientos sobrevive 5 o más años).

Dentro de las sociedades de hecho o unipersonales, se distingue luego de acuerdo al giro de actividad resultando un grupo mayoritario (con los giros C, F, G, I, J, P, R, S y T) en el que apenas 1 de cada 3 emprendimientos logra el éxito. El conjunto que no presenta estos giros, agrupa 1 de cada 4 emprendimientos, mostrando una mayor proporción de éxito de lograr la sobrevivencia.

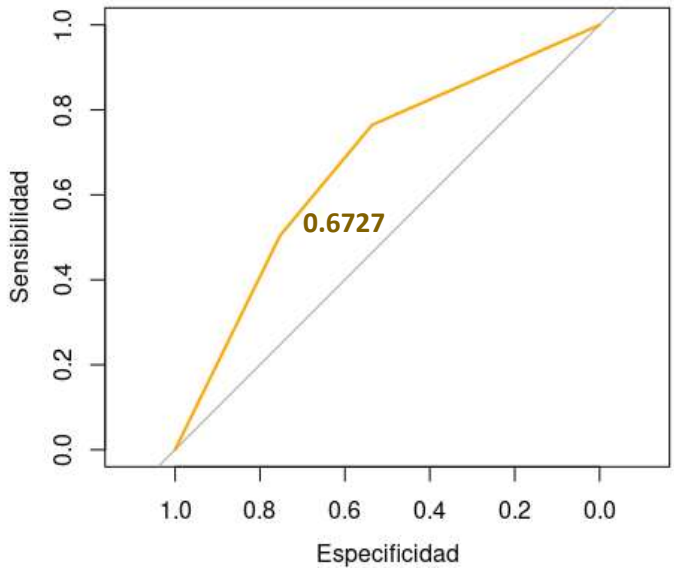
Realizada la predicción del árbol, encontramos que para este caso se obtiene un mayor valor de los coeficientes del Valor de información y de K-S, con respecto al caso del total de emprendimientos.

Cuadro 4 _ Casos presentes y ausentes en el árbol con dependientes.

Valor de información del modelo:			0,44				
Coeficiente K-S			30,05				
Nodos	Total	Ausentes	Presentes	Tot(%)	Aus(%)	Pres(%)	Prob_Pres(%)
3	1282	408	874	37.88	24.73	50.40	68.17
5	810	358	452	23.94	21.70	26.07	55.80
4	1292	884	408	38.18	53.58	23.53	31.58
Total	3384	1650	1734	100.00	100.00	100.00	no aplica

Para el cálculo de la curva Roc, el valor alcanzado del AUC es **0.6727**, por lo que también en este caso puede considerarse como regular el test.

Gráfico 6 _ Curva Roc del árbol con dependientes

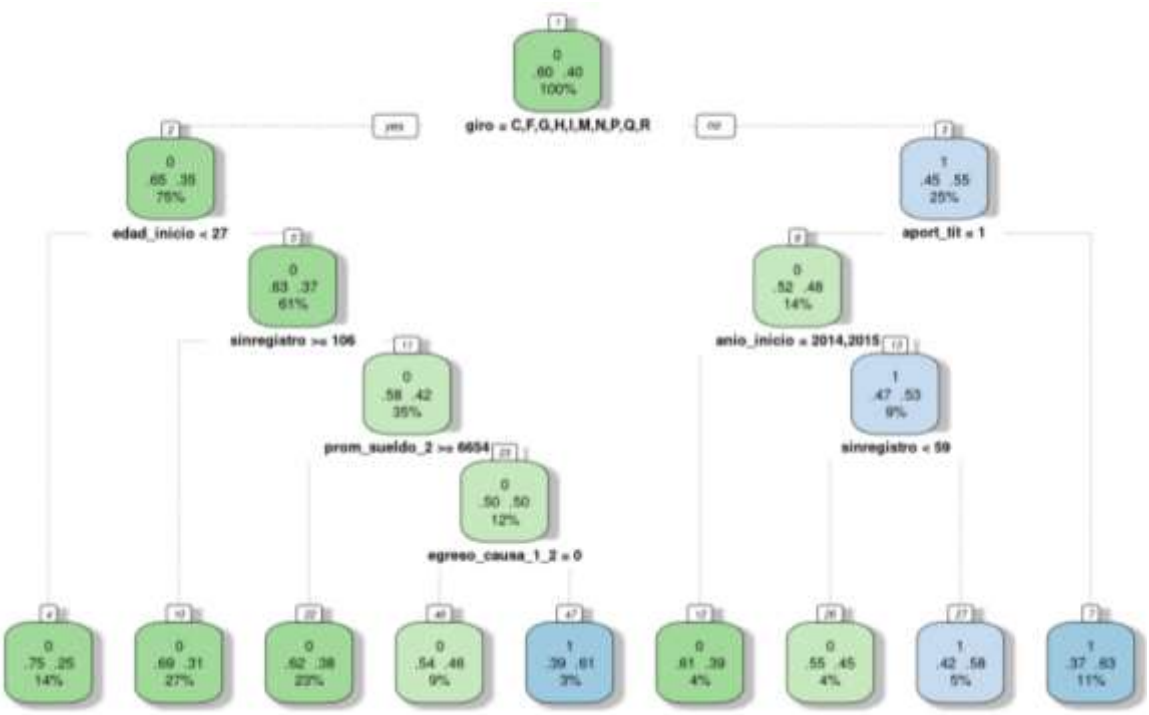


Base filtrada para los emprendimientos sin dependientes

Para la base de emprendimientos con los Patrones que no presentan trabajadores dependientes, la variable que el algoritmo selecciona como principal para distinguir los diferentes grupos es el **Giro**. La otra variable en importancia es el **Tipo de aportación**, y entran en discusión otras variables como Zona geográfica, si el emprendedor tuvo experiencia en el giro del Emprendimiento que inicia, y el año de inicio del emprendimiento.

La variable **año de inicio** entiende el algoritmo que los emprendimientos que comenzaron en los años 2014 y 2015 proyectan menor sobrevivencia, y esto puede tener al menos dos lecturas. Primero, como ya se había visto para el total, un mayor dinamismo de la economía influye positivamente en la sobrevivencia que pueda lograr un emprendimiento, y entre 2017-2019 el crecimiento de la economía uruguaya se vio ralentizado. Segundo, en 2020 el impacto en la economía que tuvo la pandemia por Covid-19 fue muy claro. Caída del PBI, aumento en la cantidad de subsidios por desempleo, cierre y caída en la actividad de las empresas.

Árbol de clasificación de la base sin dependientes



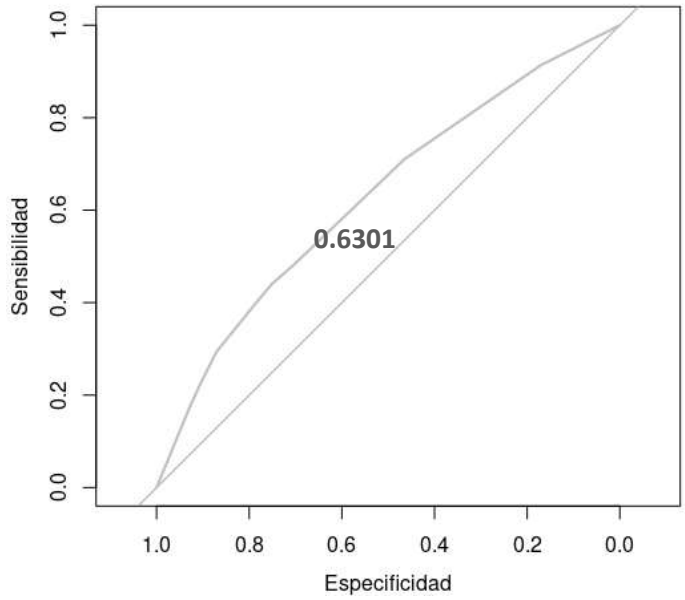
Como es de esperar, el resultado de la predicción del árbol debe arrojar valores desfavorables (respecto al caso Con dependientes). En este caso el Valor de información y el K-S son sensiblemente menores a los obtenidos con anterioridad.

Cuadro 5 _ Casos presentes y ausentes en el árbol sin dependientes.

Valor de información del modelo:			0,23				
Coeficiente K-S			19,14				
Nodos	Total	Ausentes	Presentes	Tot(%)	Aus(%)	Pres(%)	Prob_Pres(%)
7	575	214	361	11.21	7.07	17.18	62.78
47	167	66	101	3.26	2.18	4.81	60.48
27	265	111	154	5.17	3.67	7.33	58.11
46	460	247	213	8.97	8.16	10.14	46.30
26	214	117	97	4.17	3.86	4.62	45.33
12	220	137	83	4.29	4.52	3.95	37.73
22	1211	728	483	23.61	24.04	22.99	39.88
10	1321	893	428	25.76	29.49	20.37	32.40
4	696	515	181	13.57	17.01	8.61	26.01
Total	5129	3028	2101	100.00	100.00	100.00	no aplica

Para el cálculo de la curva Roc, el valor alcanzado del AUC es **0.6301**, por lo que también en este caso puede considerarse como regular el test.

Gráfico 7 _ Curva Roc del árbol sin dependientes



D _ Modelo inicial __

Para trabajar en el software R con la regresión logit previamente se realizan todas las transformaciones necesarias de las variables y se excluyen algunas dummy para evitar colinealidad.

Al comienzo, se conforma un vector de variables que no se tienen en cuenta para la regresión, en el entendido de que no presentan información relevante para el análisis. La ponderación de la base se hace de acuerdo a la variable del expansor.

Luego de obtener una primera salida de la regresión, se aplica el método “stepwise” de manera de eliminar hacia atrás los coeficientes no significativos. Posteriormente se descartan todas las variables no significativas que indica la sentencia “vif” (factor de inflación de la varianza); el mismo cuantifica la intensidad de la multicolinealidad. Las variables que den valores mayores que 2 es recomendable excluirlas de la regresión, obteniendo así una fórmula preliminar.

El proceso de selección de variables tiene el objetivo de encontrar un modelo “parsimonioso”, es decir un modelo con mayor poder explicativo pero con el menor número de variables posible.

El último filtro lo realizamos para eliminar las variables que, luego de haber realizado todos los ajustes en el proceso, no resultan ser significativas en la salida.

El modelo logit con el que trabajaremos tiene omitidas las variables dummy “aport_tit_1”, “nj_Otros”, “anio_inicio_2010”, “zg_Centro”, “giro_C” y “sexo_2”.

Corresponden a las binarias Tipo de aportación IyC, naturaleza jurídica Otros, año de inicio del emprendimiento en 2010, Centro como la zona geográfica del emprendimiento, actividad Industrias

Manufactureras y finalmente sexo del emprendedor Mujer. Las comparaciones deberán realizarse frente a estas variables.

```
Call:
glm(formula = sobreviven_T5 ~ Patr_SDep1 + prom_masasal1 + prom_puestos1 + edad_inicio + fuepatron21
+ qfuepatron + depte_t0_hila_21 + no_depte_t0_hila_21 + sinregistro + egreso_causa_1_21 +
egreso_causa_3_21 + con_desempleo_21 + con_enfermedad_21 + con_materpater_21 + qdesempleo +
prom_sueldo_2 + mismo_giro_21 + mismo_codigo_giro_21 + var_pbi_ac_12 + aport_tit_3 + nj_Cooperativa
+ nj_SH + nj_SRL + nj_Unipersonal + anio_inicio_2011 + anio_inicio_2012 + anio_inicio_2013 +
anio_inicio_2014 + anio_inicio_2015 + zg_Este + zg_Lit_Norte + zg_Lit_Sur + zg_Metropolitana +
zg_Noreste + giro_A + giro_F + giro_G + giro_H + giro_I + giro_J + giro_K + giro_L + giro_M +
giro_N + giro_P + giro_Q + giro_R + giro_S + giro_T + sexo_1,
family = "binomial", data = Base[names(Base) %nin%
    vars_excl_60], weights = Base$exp)
```

E_ Paso a paso del proceso

*Árbol de clasificación_*

Comenzamos corriendo el script para generar los árboles de clasificación con todas las variables construidas, y luego de analizar los resultados que arrojaron las primeras salidas comenzamos a tomar decisiones de selección de variables.

Seleccionamos las principales variables de la Base (25) con las que trabajaremos y también restringimos a un mínimo las cantidades en los nodos intermedios (9%) y las terminales (3%), de manera de que los grupos finales sean representativos.

```
vars_excl <- names(Base_muestra[, c(1, 24)])
set.seed(12345)
arbol.ini_1 <- rpart(sobreviven_T5 ~ .,
                     Base_muestra[ , names(Base_muestra) %nin% vars_excl],
                     method = 'class', cp = 0, weights = Base_muestra$exp, minsplit = 767, minbucket
= 256)
```

La siguiente tabla con los valores de los errores asociados a la cantidad de nodos, surge de la sentencia “printcp” en R.

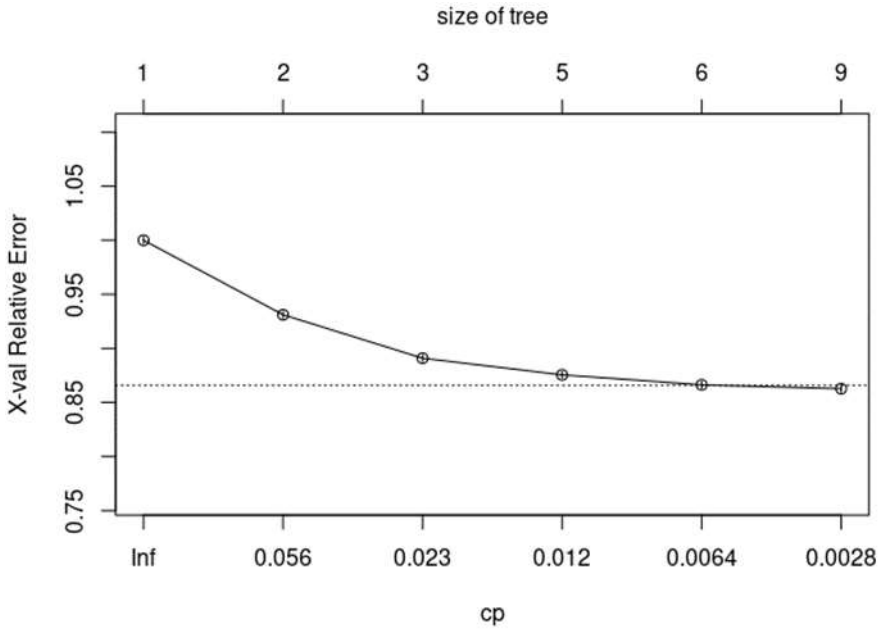
```
         CP nsplit rel error  xerror      xstd
1 0.0747957      0   1.00000 1.00000 0.0031685
2 0.0413229      1   0.92520 0.93111 0.0031362
3 0.0124794      2   0.88388 0.89092 0.0031118
4 0.0108235      4   0.85892 0.87551 0.0031014
5 0.0037745      5   0.84810 0.86630 0.0030949
6 0.0020423      8   0.83678 0.86272 0.0030923
7 0.0000000      9   0.83473 0.86274 0.0030923
```

A su vez, lo siguiente es el valor óptimo de “cp” que indica el modelo

```
         CP      nsplit   rel error      xerror        xstd
0.002042261 8.000000000 0.836775779 0.862724484 0.003092256
```

Le indicamos al árbol que tome el “mejor” cp, dado el óptimo y la variación del error según vaya generando más nodos.

```
> cp
[1] 0.002042261
```




```
Call:
rpart(formula = sobreviven_T5 ~ ., data = Base_muestra[, names(Base_muestra) %nin%
    vars_excl], weights = Base_muestra$exp, method = "class",
    cp = 0, minsplit = 767, minbucket = 256)
  n= 8513
```

```
           CP nsplit rel error    xerror        xstd
1 0.074795689      0 1.0000000 1.0000000 0.003168472
2 0.041322929      1 0.9252043 0.9311056 0.003136172
3 0.012479360      2 0.8838814 0.8909166 0.003111822
4 0.010823475      4 0.8589227 0.8755129 0.003101388
5 0.003774469      5 0.8480992 0.8663029 0.003094854
6 0.002042261      8 0.8367758 0.8627245 0.003092256
```

Luego de elegir el mejor cp:

```
Variable importance
             giro                nj         aport_tit        sinregistro
               33                21                17                 5
      edad_inicio        qfuepatron     prom_sueldo_2 no_depte_t0_hila_2
                5                 5                 4                 3
    var_pbi_ac_12                zg       anio_inicio         fuepatron2
                2                 2                 2                 1
```

Estas son las variables que el árbol considera más importantes para hacer la clasificación.

Luego se genera el árbol, que es el que está en el apartado IV.4.1, y a continuación por medio de una función se genera la tabla de performance a partir de la variable cualitativa.

```
pred_arbol.fin_1 <- predict(arbol.fin_1, Base_muestra)[, 2]

rbm <- report_cat(pred_arbol.fin_1, Base_muestra$sobreviven_T5)
```

Finalmente se elabora la curva Roc a través de las siguientes sentencias.

```
auroc <- roc(Base_muestra$sobreviven_T5, pred_arbol.fin_1)

plot(auroc, col = 'red', xlab='Especificidad', ylab='Sensibilidad')
```

La tabla y el gráfico están en el apartado IV.4.2.

*Regresión logística_*

En el Anexo D se comentó brevemente el inicio del proceso de la regresión. Recordando, luego de obtener la primera salida se realiza el "step" eliminando hacia atrás los coeficientes no significativos. Posteriormente se eliminan todas las variables que indica la sentencia "vif", y se descartan aquellas variables no significativas en la salida, llegando a la siguiente formula

```
Call:
glm(formula = sobreviven_T5 ~ Patr_SDep + prom_puestos1 + edad_inicio +
    fuepatron2 + depte_t0_hila_2 + sinregistro + egreso_causa_1_2 +
    egreso_causa_3_2 + con_enfermedad_2 + con_materpater_2 + qdesempleo +
    mismo_giro_2 + mismo_codigo_giro_2 + var_pbi_ac_12 + aport_tit_3 +
    nj_Cooperativa + nj_SH + nj_SRL + anio_inicio_2011 + anio_inicio_2013 +
    anio_inicio_2014 + zg_Este + zg_Lit_Norte + zg_Noreste + giro_F + giro_H +
    giro_I + giro_J + giro_K + giro_L + giro_M + giro_N + giro_Q + giro_S +
    giro_T + sexo_1,
    family = "binomial", data = Base[names(Base) %nin% vars_excl_60],
    weights = Base$exp)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-9.788  -4.087  -2.483   4.431   9.050

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.5765444  0.0342347 -16.841  < 2e-16 ***
Patr_SDep1           -0.2615112  0.0145019 -18.033  < 2e-16 ***
prom_puestos1         0.0205178  0.0019490  10.527  < 2e-16 ***
edad_inicio           0.0088775  0.0005333  16.647  < 2e-16 ***
fuepatron21           0.1014113  0.0163233   6.213 5.21e-10 ***
depte_t0_hila_21     -0.1047037  0.0131847  -7.941 2.00e-15 ***
sinregistro          -0.0021442  0.0001085 -19.771  < 2e-16 ***
egreso_causa_1_21    -0.1400346  0.0129505 -10.813  < 2e-16 ***
egreso_causa_3_21    -0.3110962  0.0174768 -17.801  < 2e-16 ***
con_enfermedad_21    -0.1943923  0.0196676  -9.884  < 2e-16 ***
con_materpater_21     0.1589799  0.0407677   3.900 9.63e-05 ***
qdesempleo           -0.0057363  0.0010200  -5.624 1.87e-08 ***
mismo_giro_21         0.0832582  0.0136091   6.118 9.49e-10 ***
mismo_codigo_giro_21  0.0723181  0.0184242   3.925 8.67e-05 ***
```

```
var_pbi_ac_12          0.0437279  0.0036414  12.008  < 2e-16 ***
aport_tit_3            0.8821880  0.0193968  45.481  < 2e-16 ***
nj_Cooperativa         1.2954257  0.1063888  12.176  < 2e-16 ***
nj_SH                 -0.5005465  0.0226974 -22.053  < 2e-16 ***
nj_SRL                 0.4629908  0.0222774  20.783  < 2e-16 ***
anio_inicio_2011       0.0898042  0.0175369   5.121 3.04e-07 ***
anio_inicio_2013      -0.0306949  0.0157853  -1.945   0.0518 .
anio_inicio_2014      -0.1615727  0.0169251  -9.546  < 2e-16 ***
zg_Este               -0.1168540  0.0174671  -6.690 2.23e-11 ***
zg_Lit_Norte          -0.3017790  0.0219565 -13.744  < 2e-16 ***
zg_Noreste            -0.1595533  0.0250714  -6.364 1.97e-10 ***
giro_F                -0.1629562  0.0275601  -5.913 3.36e-09 ***
giro_H                 0.4246593  0.0250371  16.961  < 2e-16 ***
giro_I                -0.5321359  0.0282506 -18.836  < 2e-16 ***
giro_J                 0.3366978  0.0356348   9.449  < 2e-16 ***
giro_K                 0.5801403  0.0719214   8.066 7.25e-16 ***
giro_L                 0.4730196  0.0479249   9.870  < 2e-16 ***
giro_M                 0.2754317  0.0248836  11.069  < 2e-16 ***
giro_N                 0.0507634  0.0258325   1.965   0.0494 *
giro_Q                 0.1462149  0.0368743   3.965 7.33e-05 ***
giro_S                 0.4801905  0.0245069  19.594  < 2e-16 ***
giro_T                -0.8915703  0.1916550  -4.652 3.29e-06 ***
sexo_1                 0.0497049  0.0124666   3.987 6.69e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 179722  on 8512  degrees of freedom
Residual deviance: 169610  on 8476  degrees of freedom
AIC: 168234

Number of Fisher Scoring iterations: 5
```

En el apartado IV.4.3 se representan por medio de 2 gráficos los coeficientes y los odds ratios que se generan a partir de la salida de la regresión.

A continuación y de manera análoga a lo realizado para el árbol, se genera la tabla de predicción de la regresión y también se elabora el gráfico de la curva Roc, ambos presentados en el apartado IV.4.4.

```
Base$pred.reg.logistic <- predict(reg.log.fin, Base, type = 'response')*100

report_cont(Base$pred.reg.logistic, Base$sobreviven_T5)

auroc <- roc(Base$sobreviven_T5, Base$pred.reg.logistic)

plot(auroc, col = 'green', xlab='Especificidad', ylab='Sensibilidad')
```