

# UNIVERSIDAD TÉCNICA DE AMBATO



## FACULTAD DE INGENIERÍA EN SISTEMAS, ELECTRÓNICA E INDUSTRIAL

### MAESTRÍA EN MATEMÁTICA APLICADA

---

**Tema:** “Modelado matemático predictivo para la sostenibilidad de los emprendimientos productivos”.

---

Trabajo de titulación previo a la obtención del grado académico de  
Magíster en Matemática Aplicada

Modalidad de titulación Proyecto de Desarrollo

**Autor:** Ing. Héctor Alberto Luzuriaga Jaramillo, Mg.

**Director:** Ing. Ernesto Antonio Ponsot Balaguer, PhD

Ambato – Ecuador

2021

## APROBACIÓN DEL TRABAJO DE TITULACIÓN

A la Unidad Académica de Titulación de la Facultad de Ingeniería en Sistemas, Electrónica e Industrial.

El Tribunal receptor del Trabajo de Investigación presidido por la Ingeniera Elsa Pilar Urrutia Urrutia Magíster, e integrado por los señores: Dr. Manuel Antonio Meneses Freire, PhD, Dr. Saba Rafael Infante Quirpa, PhD, designados por la Unidad Académica de Titulación de la Facultad de Ingeniería en Sistemas, Electrónica e Industrial de la Universidad Técnica de Ambato, para receptor el Trabajo de Investigación con el tema: “Modelado matemático predictivo para la sostenibilidad de los emprendimientos productivos”, elaborado y presentado por el señor Ingeniero Héctor Alberto Luzuriaga Jaramillo, Magister, para optar por el Grado Académico de Magíster en Matemática Aplicada; una vez escuchada la defensa oral del Trabajo de Investigación el Tribunal aprueba y remite el trabajo para uso y custodia en las bibliotecas de la UTA.



---

Ing. Elsa Pilar Urrutia Urrutia, Mg.  
Presidente del Tribunal



---

Dr. Manuel Antonio Meneses Freire, PhD.  
Miembro del Tribunal



---

Dr. Saba Rafael Infante Quirpa, PhD.  
Miembro del Tribunal

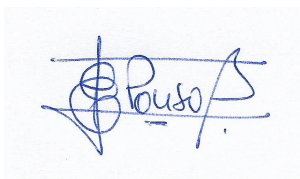
## AUTORÍA DEL TRABAJO DE TITULACIÓN

La responsabilidad de las opiniones, comentarios y críticas emitidas en el Trabajo de Titulación presentado con el tema: “Modelado matemático predictivo para la sostenibilidad de los emprendimientos productivos”, le corresponde exclusivamente a: Ingeniero Héctor Alberto Luzuriaga Jaramillo, Magister, Autor bajo la Dirección del Ingeniero Ernesto Antonio Ponsot Balaguer, PhD, Director del Trabajo de Investigación; y el patrimonio intelectual a la Universidad Técnica de Ambato.



---

Ing. Héctor Alberto Luzuriaga Jaramillo, Mg.  
**AUTOR**



---

Ing. Ernesto Antonio Ponsot Balaguer, PhD  
**DIRECTOR**

## DERECHOS DE AUTOR

Autorizo a la Universidad Técnica de Ambato, para que el Trabajo de Titulación, sirva como un documento disponible para su lectura, consulta y procesos de investigación, según las normas de la Institución.

Cedo los Derechos de mi Trabajo de Titulación, con fines de difusión pública, además apruebo la reproducción de este, dentro de las regulaciones de la Universidad Técnica de Ambato.



---

Ing. Héctor Alberto Luzuriaga Jaramillo, Mg.

cc. 1803253150

**AUTOR**

# Índice general

Índice de tablas	vii
Índice de figuras	viii
Resumen	xii
Executive Summary	xiv
<b>1. El problema de investigación</b>	<b>1</b>
1.1. Introducción . . . . .	1
1.2. Justificación . . . . .	2
1.3. Objetivos . . . . .	3
1.3.1. Objetivo general . . . . .	3
1.3.2. Objetivos específicos . . . . .	3
<b>2. Antecedentes investigativos</b>	<b>4</b>
2.1. Estado del arte . . . . .	4
2.2. Marco teórico . . . . .	12
2.2.1. El modelo matemático . . . . .	12
2.2.2. El modelo estadístico . . . . .	12
2.2.3. La familia exponencial de distribuciones . . . . .	14
2.2.4. Modelo lineal generalizado (MLG) . . . . .	15
2.2.5. La función de masas binomial . . . . .	16
2.2.6. El modelo de regresión logística . . . . .	16
2.2.7. Pruebas de hipótesis e intervalos de confianza sobre $\hat{\beta}$ . . . . .	19
2.2.8. Pruebas de bondad del ajuste . . . . .	20
2.2.9. Técnicas de selección de variables . . . . .	22
2.2.10. Curvas ROC . . . . .	22
2.2.11. Emprendimientos . . . . .	24
<b>3. Marco metodológico</b>	<b>26</b>
3.1. Ubicación . . . . .	26
3.2. Equipos y materiales . . . . .	26
3.3. Tipo de investigación . . . . .	27
3.4. Pruebas de Hipótesis . . . . .	27
3.4.1. Preguntas que motivan la investigación . . . . .	27
3.4.2. Sistema de Hipótesis . . . . .	28
3.5. Población y muestra . . . . .	28
3.5.1. Recolección de información . . . . .	29

3.6. Procesamiento de la información y análisis estadístico . . . . .	29
<b>4. Resultados y discusión</b>	<b>31</b>
4.1. Análisis Descriptivo de los datos . . . . .	31
4.1.1. Análisis unidimensional . . . . .	31
4.1.2. Análisis bidimensional . . . . .	41
4.2. El modelo de regresión logística . . . . .	45
4.2.1. Las tablas ANOVA . . . . .	47
4.2.2. El modelo seleccionado . . . . .	48
4.3. El modelo seleccionado como predictor / clasificador . . . . .	56
<b>5. Conclusiones y recomendaciones</b>	<b>58</b>
5.1. Conclusiones . . . . .	58
5.2. Recomendaciones . . . . .	60
<b>Bibliografía</b>	<b>61</b>
<b>Anexos</b>	<b>64</b>
Anexo 1: Instrumento de recolección de datos . . . . .	64
Anexo 2: Código R programado . . . . .	69

# Índice de tablas

2.1. Matriz de confusión . . . . .	23
3.1. Población y muestra por cantones . . . . .	29
3.2. Variables y tipos . . . . .	30
4.1. Distribución de frecuencias para la variable Género . . . . .	31
4.2. Distribución de frecuencias para la variable Edad . . . . .	32
4.3. Distribución de frecuencias para la variable Instrucción . . . . .	33
4.4. Distribución de frecuencias para la variable Presupuesto . . . . .	33
4.5. Distribución de frecuencias para la variable Tiempo_emprende . . . . .	34
4.6. Distribución de frecuencias para la variable Actividad . . . . .	37
4.7. Distribución de frecuencias para la variable Aspectos_emprende . . . . .	37
4.8. Distribución de frecuencias para la variable Problemas . . . . .	38
4.9. Distribución de frecuencias para la variable Publicidad . . . . .	39
4.10. Distribución de frecuencias para la variable Personal . . . . .	40
4.11. Distribución de frecuencias para la variable Cantón . . . . .	41
4.12. Distribución de frecuencias para la variable Suspendido . . . . .	42
4.13. Distribución de frecuencias para la variable Cerrado . . . . .	43
4.14. Distribución conjunta por edad y género . . . . .	44
4.15. Distribución conjunta por Edad e Instrucción . . . . .	44
4.16. Distribución conjunta por Edad y Cantón . . . . .	45
4.17. Correspondencia entre las variables del archivo de datos y las del modelo . . . . .	48
4.18. Análisis de varianza . . . . .	49
4.19. Análisis de varianza (Cont.) . . . . .	50
4.20. Análisis de varianza para el modelo seleccionado . . . . .	51
4.21. Coeficientes e intervalos de confianza al 90 % . . . . .	53
4.22. Matrices de confusión (%) para $p_0 = 0,5; 0,4; 0,3; 0,2$ . . . . .	57

# Índice de figuras

2.1.	Diagrama de dispersión para $x$ e $y$ y recta de regresión. . . . .	14
2.2.	Factores del éxito empresarial . . . . .	24
2.3.	Evolución de la Actividad Emprendedora Temprana en Ecuador y países seleccionados . . . . .	25
4.1.	Distribución por género. . . . .	32
4.2.	Distribución por edad. . . . .	33
4.3.	Nivel de instrucción de los emprendedores. . . . .	34
4.4.	Presupuesto promedio anual de los emprendedores. . . . .	35
4.5.	Tiempo que lleva el emprendimiento. . . . .	36
4.6.	Aspectos considerados para el emprendimiento. . . . .	38
4.7.	Problemas para el desarrollo del emprendimiento. . . . .	39
4.8.	Tipo de publicidad para el emprendimiento. . . . .	40
4.9.	Personal de apoyo en relación de dependencia. . . . .	41
4.10.	Cantones objeto de la investigación. . . . .	42
4.11.	Emprendimientos suspendidos. . . . .	43
4.12.	Emprendimientos cerrados. . . . .	43
4.13.	Diagrama de caja de la edad y género. . . . .	45
4.14.	Diagrama de caja de la edad y el nivel de instrucción. . . . .	46
4.15.	Diagrama de caja de la edad y el cantón. . . . .	47
4.16.	Curva ROC del modelo seleccionado ( $p_0 = 0,3$ ). . . . .	57
1.	Instrumento de recolección de datos. Página 1. . . . .	64
2.	Instrumento de recolección de datos. Página 2. . . . .	65
3.	Instrumento de recolección de datos. Página 3. . . . .	66
4.	Instrumento de recolección de datos. Página 4. . . . .	67
5.	Instrumento de recolección de datos. Página 5. . . . .	68



# Agradecimiento

Por sobre todas las cosas a Dios.  
A mis padres que me enseñaron que  
nada es inalcanzable en este mundo.  
A mi esposa Andreita, toda su comprensión  
y apoyo en todo momento.  
Al tutor de este trabajo por su  
paciencia en la enseñanza.  
**Alberto Luzuriaga**

# Dedicatoria

*El presente trabajo se lo  
dedico en primer lugar a Dios  
y a toda mi familia, en especial  
a mi esposa Andreita, a mi chiquito  
son mi apoyo incondicional.*

**Alberto Luzuriaga**

**UNIVERSIDAD TÉCNICA DE AMBATO**

**FACULTAD DE INGENIERÍA EN SISTEMAS, ELECTRÓNICA  
E INDUSTRIAL**

**MAESTRÍA EN MATEMÁTICA APLICADA**

**TEMA:**

**“MODELADO MATEMÁTICO PREDICTIVO PARA LA SOSTENIBILIDAD  
DE LOS EMPRENDIMIENTOS PRODUCTIVOS”.**

**AUTOR:** Ing. Héctor Alberto Luzuriaga Jaramillo, Mg.

**DIRECTOR:** Ing. Ernesto Antonio Ponsot Balaguer, PhD

**LÍNEA DE INVESTIGACIÓN:**

- Diseño, Materiales y Producción.
- Tecnología de la información.

**FECHA:** 4 de enero de 2021

# Resumen

Los modelos que explican y predicen el fracaso y el éxito son muy valiosos para los emprendedores del sector de la producción. Se presenta la implementación de un modelo matemático - estadístico explicativo y predictivo, para el estudio de la sostenibilidad de los emprendimientos productivos en los cantones Cevallos, Quero, Mocha, Patate y Tisaleo de la provincia de Tungurahua, Ecuador. Se propone y ajusta un modelo de regresión logística sobre una muestra de 1546 emprendedores. El método de selección de variables es *stepwise* que, siguiendo el principio de parsimonía, incorpora solamente las variables que resulten significativas. La prueba de bondad del ajuste del deviance sugiere que dicho modelo ajusta apropiadamente los datos. Se comprueban y comentan varias hipótesis, siendo el resultado más importante, la demostración de que existen factores que pueden ser utilizados para explicar el éxito de los emprendimientos productivos. Estos son: instrucción, tiempo del emprendimiento, personal, cantón y si ha sido suspendido.

**Palabras clave:** Análisis de varianza, Deviance, Emprendimientos productivos, Encuestas, Logit, Modelo lineal generalizado, Predicción, Regresión logística, Resultados empresariales, ROC.

**UNIVERSIDAD TÉCNICA DE AMBATO**

**FACULTAD DE INGENIERÍA EN SISTEMAS, ELECTRÓNICA  
E INDUSTRIAL**

**MAESTRÍA EN MATEMÁTICA APLICADA**

**THEME:**

**“PREDICTIVE MATHEMATICAL MODELING FOR THE  
SUSTAINABILITY OF PRODUCTIVE ENTREPRENEURSHIPS”.**

**AUTHOR:** Ing. Héctor Alberto Luzuriaga Jaramillo, Mg.

**DIRECTED BY:** Ing. Ernesto Antonio Ponsot Balaguer, PhD

**LINE OF RESEARCH:**

- Design, Materials and Production.
- Information Technology.

**DATE:** January 4, 2021

# Executive Summary

Models that explain and predict failure and success are very valuable to entrepreneurs in the production sector. The implementation of an explanatory and predictive mathematical - statistical model is presented for the study of the sustainability of productive enterprises in the Cevallos, Quero, Mocha, Patate and Tisaleo cantons of the province of Tungurahua, Ecuador. A logistic regression model is proposed and adjusted on a sample of 1546 entrepreneurs. The variable selection method is stepwise which, following the parsimonious principle, incorporates only the variables that are significant. The deviance goodness-of-fit test suggests that such a model fits the data appropriately. Several hypotheses are checked and discussed, the most important result being the demonstration that there are factors that can be used to explain the success of productive entrepreneurs. These are: Instruction, time of the undertaking, staff, canton and if it has been suspended.

**Keywords:** Analysis of variance, Deviance, Productive entrepreneurships, Surveys, Logit, Generalized linear model, Prediction, Logistic regression, Business results, ROC.

# Capítulo 1

## El problema de investigación

### 1.1. Introducción

Se estudia la sostenibilidad de los emprendimientos productivos en función de algunas características, entendidas como variables explicativas. Estas son objeto de análisis mediante un modelo matemático predictivo (o estadístico), para relacionar el éxito y el fracaso de dichos emprendimientos.

Los datos se toman del estudio de sostenibilidad de los emprendimientos productivos de la Provincia de Tungurahua, Ecuador, Frente 1 - Fase 2, desarrollado por la Facultad de Contabilidad y Auditoría de la Universidad Técnica de Ambato, específicamente en los cantones Cevallos, Quero, Mocha, Patate y Tisaleo.

En principio se realiza un análisis de la problemática de los emprendimientos productivos y se postulan interrogantes, que dan lugar a las hipótesis de trabajo. Se justifica el tema y se propone el objetivo perseguido.

A continuación, se identifica la base científica y el fundamento teórico que sustenta el trabajo, con la reseña de diversos artículos científicos de alto impacto, que tienen estrecha relación con las variables de estudio.

Se utiliza un paradigma crítico propositivo con un enfoque cuantitativo, sobre una muestra de 1546 emprendimientos. El modelo estadístico utilizado es la Regresión Logística, implementación particular del Modelo Lineal Generalizado.

Se realiza un análisis descriptivo unidimensional y bidimensional, que incluye la interpretación de los resultados obtenidos para las 12 variables explicativas y la variable de respuesta. Un total de 13 variables fueron seleccionadas de 28 variables que contenía el estudio original (ver Anexo 1).

En la última sección, la parte medular del trabajo, se busca el mejor modelo de regresión logística. Esto permite comprobar las hipótesis presentadas en el documento. Además, se analiza la capacidad de predicción del modelo, empleando la curva ROC (por las siglas en inglés de *Receiver Operating Characteristic*).

El desarrollo se limita a los cinco cantones mencionados de la Provincia de Tungurahua ya que, debido a la situación de emergencia sanitaria en el país, no se pudo aplicar el instrumento en todos los cantones de la provincia, como era el objetivo inicial. En tal sentido, se trata de una base para futuras investigaciones.

## 1.2. Justificación

El presente trabajo tiene el potencial de generar un impacto significativo en el sector productivo, en la búsqueda del mejor rendimiento de los capitales y una adecuada toma de decisiones de inversión.

Los beneficiarios directos del análisis son los productores y empresarios, que disponen de información explicativa y de carácter predictivo, con la que pueden anticiparse a cierres o suspensiones de sus emprendimientos, actuando a tiempo para evitarlos.

El tema se torna novedoso debido al proceso tecnológico empleado. Se trata de una metodología basada en el aprendizaje estadístico, que incluye modelos programados en software especializado. Además, utiliza gran cantidad de variables explicativas de entrada, para describir la sostenibilidad de los emprendimientos.

Se trata de un aporte original, pues no se tienen noticias de que se hayan desarrollado investigaciones de corte similar en el país. Sólo se ha encontrado una investigación sobre las causas de la mortalidad de compañías en la región 3 en el Ecuador, según Núñez, Peñalosa, y Armijos (2015), pero el tema se expone sólo teóricamente.

Los resultados de la presente investigación se podrán utilizar como un estado del arte para futuros proyectos y sus resultados serán difundidos en una revista de alto impacto.



## **1.3. Objetivos**

### **1.3.1. Objetivo general**

Implementar un modelo matemático explicativo y predictivo para el estudio de la sostenibilidad de los emprendimientos productivos en los cantones Cevallos, Quero, Mocha, Patate y Tisaleo de la provincia de Tungurahua, Ecuador.

### **1.3.2. Objetivos específicos**

- Determinar las variables más significativas que sirvan de entrada al modelo matemático.
- Analizar las variables más significativas para aumentar la precisión de la expectativa de sostenibilidad de los emprendimientos productivos, en los cantones Cevallos, Quero, Mocha, Patate y Tisaleo de la provincia de Tungurahua, Ecuador.
- Proponer, ajustar y describir un modelo matemático explicativo y predictivo de la sostenibilidad de los emprendimientos productivos de los cantones Cevallos, Quero, Mocha, Patate y Tisaleo de la provincia de Tungurahua, Ecuador.

# Capítulo 2

## Antecedentes investigativos

### 2.1. Estado del arte

En el Ecuador, según Núñez y cols. (2015), el elemento principal que explica el éxito o el fracaso de los emprendimientos productivos, es la toma de decisiones de los empresarios y sus socios en base a elementos financieros, jurídicos y administrativos. Se concluye que el incumplimiento de la ley, la gestión administrativa y contable, son factores que explican el fracaso.

Esta investigación es una base para el presente trabajo, por su metodología de carácter exploratorio y descriptivo solamente, es decir, su principal limitación es no haber producido y estudiado un modelo matemático o estadístico que les permitiera predecir las fortalezas y debilidades de las causas de la mortalidad de las empresas. No obstante, los datos son muy útiles como referencia para el modelado matemático, en cuanto a la selección de las diferentes variables que pudieran explicar el problema.

Messina y Hochsztain (2015) analizan dos categorías: éxito o fracaso, aplicando técnicas de clasificación, como son los árboles de decisión y la regresión logística. Se pronostica el éxito de un emprendimiento para identificar emprendimientos potenciales y emprendimientos que necesiten apoyo de diferentes entidades. En su estudio cuantitativo realizado a emprendedores de la República del Uruguay se pueden encontrar coincidencias con los resultados obtenidos en la presente investigación.

Este trabajo demuestra que es de gran importancia el uso de técnicas de clasificación, sin embargo, tiene limitantes debido a que utiliza pocas variables. Esto

debido a la ausencia de algunas preguntas interesantes en la encuesta aplicada. También resulta limitante su diseño totalmente cuantitativo, dejando de lado variables explicativas que pueden generar información valiosa, como el ciclo de vida del emprendedor, por ejemplo, entre otras.

Lussier y Pfeifer (2001) mencionan que existen grandes diferencias entre países, en lo que a los emprendimientos productivos se refiere. El modelo reducido predice significativamente el éxito y el fracaso de las empresas fuera de los EEUU. La investigación utiliza el modelo de predicción de éxito de Lussier de 1995, desarrollado con datos de los EEUU. Se aplica a una muestra de empresas de Europa Central y Oriental, utilizando dos modelos matemáticos. Los factores encontrados como predictores en las dos investigaciones, coinciden en que los recursos humanos tienen relación con el éxito y fracaso empresarial, sin embargo, la conclusión cambia al aplicar un nuevo modelo reducido.

Considerando esta investigación se observa que es muy importante aplicar por lo menos dos modelos, tanto el modelo “completo” en el que se incluye 15 variables explicativas, como el modelo “reducido” en el que se toman en cuenta sólo las variables más significativas. Este último modelo conduce a conclusiones diferentes a las del modelo completo, poniendo como factores principales a la educación y al personal de apoyo, que predicen significativamente el éxito o el fracaso empresarial. Además, es de mucho interés por la aplicación de modelos de regresión logística, que permiten probar con todas las variables disponibles.

Navarrete y Sansore (2011) identifican todos los factores que originan el fracaso empresarial de 397 pequeñas y medianas empresas (PYMES) establecidas, aplicando un análisis econométrico en base a modelos de regresión múltiple, con mínimos cuadrados ordinarios. Para encontrar los componentes en el fracaso de las empresas, plantean el siguiente modelo:

$$y = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_4x_4 + \epsilon$$

siendo la variable  $y$  la utilidad,  $x$  las variables regresoras y  $\epsilon$  que corresponde a los residuos.

Este modelo es muy útil al momento de decidir la solución a emplear, porque si bien es cierto que no utiliza variables explicativas cualitativas (sólo regresores), sí utiliza una representación de menor dimensión, en donde se analizan las relaciones de un conjunto de  $n$  variables correlacionadas, aplicando rotación ortogonal en el espacio  $n$ -dimensional denominada rotación varimax<sup>1</sup>.

---

<sup>1</sup>Rotación ortogonal que minimiza el número de variables que tienen cargas altas en cada

La investigación identifica los diferentes factores que condicionan la sostenibilidad en el mercado y sugiere cómo implementar políticas públicas de gobierno. Sin embargo, tiene limitantes porque relaciona la economía del Estado solo con servicios y comercio. Esto deja abierta la posibilidad de estudios posteriores para ratificar o rectificar sus afirmaciones, con otros datos.

Halabi y Lussier (2014) postulan un modelo probit de tipo ordenado que explica y predice el desempeño relativo de las pequeñas empresas en Chile, América del Sur. Las empresas fueron clasificadas como 158 fallidas, 101 mediocres y 144 exitosas. Incluye pruebas de análisis de variación (ANOVA, por sus siglas en inglés) con tres niveles de desempeño. El estudio es más robusto que el de Lussier que realiza regresión logística con sólo dos niveles de desempeño. Concluyen que el modelo utilizado se puede aplicar en diferentes campos, específicamente en temas económicos, con el objetivo de perfeccionar la probabilidad de éxito. Por lo tanto, un emprendedor, empresario, gerente o personas que tienen una idea de negocios, puedan utilizar el modelo para percibir, corregir o comprobar de mejor manera qué tipo de características o factores son necesarios para aumentar la probabilidad de éxito.

Esta investigación presenta puntos fundamentales para mejorar la precisión en el modelo de la probabilidad de éxito. Aunque el análisis no es específico para los temas de emprendimiento a nivel productivo o comercial o de otro sector económico, el modelo es muy fiable porque predice el rendimiento relativo, faltaría complementar o rectificar, en futuros estudios, las características distintas como el aspecto contable de un emprendimiento.

Otra de las técnicas que son de actualidad es la aplicación de *Machine Learning* para predecir el éxito o el fracaso. Por ejemplo, en el estudio de Yeh, Chi, y Hsu (2010) describen la importancia de las predicciones en el fracaso empresarial, aplicando una combinación de modelos que ajustan de mejor manera la predicción. Utilizan variables predictivas integradas en una técnica de teoría de conjuntos aproximados y máquinas de soporte vectorial, lo que ayuda a la simplificación de variables independientes. Luego aplican un enfoque de redes neuronales de propagación hacia atrás.

Los modelos híbridos se complementan de manera que la precisión aumenta en la predicción del fracaso empresarial, por lo que es interesante el uso de *Machine Learning* como una variante que puede utilizarse al disponer de una gran cantidad de datos.

---

factor.

El estudio desarrollado por Ponsot y Sinha (2009) formula un modelo lineal generalizado del tipo regresión logística con probabilidades proporcionales. Identifican variables que explican el resultado de los estudios universitarios en la FACES-ULA. Además, establecen un modelo lineal generalizado del tipo regresión logística multinomial con otras variables, que explican el promedio de calificaciones de los estudios secundarios de un alumno de la FACES-ULA, cada análisis es medido por variables de manera separada.

El estudio considera a 287 estudiantes seleccionados del sistema integrado de la Universidad. Su conclusión principal es que los modelos de regresión logística demuestran ser una herramienta muy poderosa para indagar sobre la explicación de variables categóricas de respuesta, en las que el investigador tenga especial interés.

Para este trabajo resulta de mucho interés identificar estudios que involucren variables categóricas de respuesta, pues permiten indagar relaciones de causalidad. El estudio tiene un alcance importante debido al manejo estadístico, especialmente por los dos métodos utilizados: regresión logística con probabilidades proporcionales y multinomial, lo que permite mejorar la toma de decisiones. Será una base para futuros proyectos.

En el estudio de Heredia, Rodríguez, y Vilalta (2014) se sostiene que las estimaciones de probabilidad, encontradas con el modelo utilizado, son útiles para el desarrollo de un método que permite predecir el estado en la asignatura de los alumnos, es decir, trata sobre el índice de fracaso en las asignaturas.

El estudio se basa en una muestra de 274 estudiantes de cursos académicos diferentes y su principal conclusión es que la utilización de herramientas matemáticas, que apoyan la planificación metodológica en materias donde la tasa de reprobación no es despreciable, constituye una importante vía para elevar las posibilidades de éxito de los alumnos.

Con relación al estudio mencionado, se identifica que el uso de variables regresoras muestran relaciones muy significativas con la variable dependiente. Esto comprueba que el uso de la regresión logística tiene aplicación en varios campos. El estudio es una referencia para predecir una respuesta categórica u ordinal. La modalidad ordinal supone la categoría equivalente a un fracaso a la calificación más baja, hasta la calificación más alta, que determina el éxito.

Para el presente estudio de los emprendimientos productivos se analizan los éxitos y fracasos directamente, sin embargo, es de mucha utilidad conocer enfoques

diferentes. En este caso, la muestra del estudio es pequeña, pero la técnica prueba su utilidad.

Con relación a los emprendimientos, en el estudio de Alvarez, C. and Noguera, M. and Urbano, D. (2012) aplican modelos de regresión logística para analizar los factores del entorno que influyen en la probabilidad de que las mujeres sean emprendedoras en el ámbito de España. El estudio se basa en los datos del proyecto *Global Entrepreneurship Monitor* (GEM). Su resultado principal es que la percepción de habilidades para emprender, redes sociales y rol familiar, son más relevantes para el emprendimiento femenino, que los factores formales como financiación, políticas de apoyo no económicas y nivel de educación.

El estudio es relevante por la actualidad del tema, que sugiere la investigación de los emprendimientos, en futuras líneas de investigación. Además, se utilizan pocas variables explicativas, que se obtienen de un análisis preliminar que se puede complementar con varias técnicas para mejorar la eficiencia de sus predicciones. Uno de los limitantes a considerar es la ausencia de un análisis más general, referido a las hipótesis planteadas, no sólo en cuanto a las emprendedoras, sino en sentido más amplio.

Igualmente se dispone de un estudio que identifica la probabilidad del éxito al iniciar un emprendimiento (Van Gelderen, Thurik, y Bosma, 2005). Los autores estiman la importancia relativa de una variedad de enfoques y variables para explicar el éxito, previo al inicio del emprendimiento. Una vez más utilizan análisis de regresión logística sobre una muestra, así como sobre subgrupos dentro de la muestra, en este caso, sobre aquellos grupos con alta ambición, frente a los de baja ambición. También sobre aquellos con experiencia sustancial, frente a aquellos de experiencia limitada. La muestra consta de 517 emprendedores incipientes. Una vez revisados los resultados de la regresión logística, una de sus conclusiones es que una empresa a tiempo parcial puede ser una desventaja, porque el empresario se distrae de sus objetivos si no trabaja a tiempo completo. La regresión logística empleada utiliza variables categóricas ordinales y una respuesta también categórica.

Mavi, Mavi, y Goh (2015) describen el desarrollo de un modelo ANFIS (por las siglas en inglés de *Adaptive Network-based in Fuzzy Inference Systems*) que resulta clave para pronosticar el éxito del emprendimiento empresarial en organizaciones industriales. Los autores utilizan un enfoque neuro-difuso con factores críticos de éxito (motivaciones) y factores críticos de fracaso (barreras). Se basan en los datos de 464 graduados de MBA (por las siglas en inglés de *Master of Business*

*Administration*) que trabajan en organizaciones industriales. Los datos se dividen en dos grupos (entrenamiento y verificación) es decir, usan el enfoque de aprendizaje automático con modelos de redes neuronales. En una de sus interesantes conclusiones afirman que existen modelos de regresión logística para los datos disponibles, pero que sería importante adaptarlos a un proceso de aprendizaje en el ajuste de la ecuación de regresión, para mejorar el desempeño.

Kilkenny, Nalbarte, y Besser (2013) dan a conocer una prueba empírica de la importancia del apoyo comunitario recíproco, en contraste con los factores económicos tradicionales y el apoyo unilateral, en el éxito de las pequeñas empresas de las ciudades más pequeñas. La hipótesis de los autores es que los emprendedores que realizan contribuciones no comerciales a su comunidad y cuya comunidad los apoya, tienen más probabilidad de que sus negocios sean exitosos. El método utilizado es la regresión logística basado en más de 800 pequeñas empresas en 30 pequeñas ciudades del estado de Iowa (EEUU). Además, concluyen que los emprendedores son una parte activa de una comunidad que brinda apoyo mutuo, simétrico y recíproco, y que el efecto de interacción del servicio de un emprendedor con la comunidad, correspondido por el apoyo de la comunidad a la empresa, es el determinante más significativo del éxito empresarial.

Este estudio se basa en apoyos de propietarios de negocios a la comunidad y viceversa. Esto es de interés directo para el presente trabajo por la metodología utilizada (la regresión logística), y de interés futuro, al sugerir la intervención del capital social en el desarrollo regional.

Algo semejante ocurre con la investigación de Hyder y Lussier (2016). Los autores sostienen que las pequeñas y medianas empresas en Pakistan, son esenciales para la estabilidad económica del país y que actúan como un colchón para los momentos duros a nivel económico. Su metodología se basa en tres niveles de análisis estadístico, a saber: el análisis descriptivo, las pruebas t para comparar las diferencias entre 18 variables y el análisis de regresión logística, utilizando el éxito / fracaso como variable dependiente, con datos de entrevistas personales que aplicaron el cuestionario de estudio de la investigación de Lussier (1995).

El estudio se relaciona en gran medida con la presente investigación, por lo que es de gran ayuda para complementar procesos aún considerados. Por ejemplo, al disponer de más datos, se podría dar un enfoque diferente, también es importante considerar el éxito de una empresa cuando ésta ha producido utilidades y fracaso en caso contrario. En la presente investigación, sin embargo, el cierre definitivo o no del negocio son considerados fracaso y éxito, respectivamente.

Romero, Haro, y Ramírez (2017) reportan un estudio de carácter empírico basado en 2119 iniciativas emprendedoras financiadas entre 2006 y 2013. Se mide el éxito o fracaso usando variables explicativas. En principio se aplica análisis descriptivo para luego aplicar el modelo de regresión logística. Además, concluyen los autores que las variables gestión de empleo, gestión presupuestal y gestión en las ventas, son las determinantes del éxito, mientras que el fracaso está asociado a ofrecer más financiación a cambio de más empleos.

Es un estudio muy similar y de gran aporte para el presente proyecto. Su análisis se basa en gran cantidad de datos, lo que da a sus resultados un buen respaldo.

Por lo que se refiere a factores de éxito, según Messina y Hochsztain (2015) es posible identificar los principales factores asociados al éxito de un emprendimiento y cómo se vinculan para anticipar su futuro. Su metodología está basada en dos técnicas de clasificación, como son: árboles de decisión y regresión logística, ambas con resultados similares. Su principal conclusión es la importancia del financiamiento a emprendedores y de la situación laboral preexistente del emprendedor. Su limitante es que se trabaja con métodos cuantitativos de investigación, usando pocas observaciones, dejando de lado técnicas cualitativas que podrían aportar.

Otro rasgo de importancia según Morales, Bustamante, Vargas, Pérez, y Sereno (2015) es la motivación para el emprendimiento. Los autores estudian los factores del éxito competitivo de los emprendedores de pequeños negocios, que se ubican en la Montaña de Guerrero, México. Su estudio está basado en 40 microempresas en Metlatónoc y 50 en Huamuxtitlán, en las que intervienen variables sociodemográficas, estructurales y económicas para iniciar un emprendimiento. Describen los datos, aplican una prueba U de Mann-Whitney (prueba no paramétrica de heterogeneidad de dos muestras) y utilizan el software SPSS para la consolidación de sus resultados, mediante una regresión logística binaria. Finalmente concluyen que el factor psicosocial tiene que ver con la confianza del emprendedor en el éxito del negocio.

En el estudio de Gómez, Lafuente, Vaillant, y Gómez (2015) se determina que la autoconfianza en las habilidades emprendedoras y algunas otras variables socio-culturales, ejercen una influencia determinante sobre la actividad emprendedora de los jóvenes en España. En su metodología disponen de una base de datos GEM-España de 2009 con 24099 personas y aplican un modelo de regresión logística. El resultado principal de su trabajo es que los jóvenes son más emprendedores que los mayores.

Lo más importante del estudio es que realiza el análisis sobre un gran volumen



de datos. Adicionalmente sirve como referencia para la presente investigación, al momento de establecer las hipótesis y decidir qué variables tendrían mayor incidencia.

En cuanto a técnicas de regresión logística, según Silva, Freire, Lina, Tavares, y Rojas (2016) éstas verifican una diferencia de potencial emprendedor entre emprendedores exitosos y emprendedores que fracasaron. Encuentran que existen variables que pueden funcionar como medio de predicción del éxito o el fracaso de un emprendedor. En su metodología utilizan un enfoque descriptivo y cuantitativo, aplicados a 246 emprendedores utilizando, además de la regresión logística, la prueba t de Student. El estudio concluye señalando la eficacia de predicción de la escala de potencial emprendedor, para determinar la diferencia entre emprendedores exitosos y emprendedores que quebraron.

Leary y DeVaughn (2009) identifican las características de un equipo empresarial, que influyen en la probabilidad de que una nueva empresa se lance con éxito. En su metodología, nuevamente se utiliza la regresión logística para probar las hipótesis, con una muestra de bancos de reciente creación que solicitaron *charter* en Florida entre 1996 y 2005. Su principal resultado muestra que sus variables de control, el capital inicial y el tamaño del equipo, son estadísticamente significativos y están relacionados positivamente con el lanzamiento.

Estos resultados son una referencia importante, pues resaltan la significación de las características del equipo emprendedor con relación a que una nueva empresa tenga éxito o no. También representa una base preliminar para futuras investigaciones.

Por último, Laitinen (2013) destaca las predicciones sobre los fracasos de pequeñas empresas reorganizadas de acuerdo con la Ley de Reorganización de Empresas de Finlandia (FCRA). En otras palabras, reporta sobre el éxito o fracaso en el proceso de reorganización. Una vez más utiliza la regresión logística binaria. Su principal hipótesis es que las acciones de reorganización orientadas a la eficiencia, son factores críticos de éxito para disminuir el riesgo de fracaso de la reorganización.

El estudio utiliza diferentes modelos. Parte de un análisis descriptivo, para luego comparar modelos buscando un resultado más eficiente. El autor informa que el modelo de regresión logística mostró una gran ventaja en la precisión de la clasificación, sobre otra técnica comparada, denominada análisis de supervivencia.

## 2.2. Marco teórico

### 2.2.1. El modelo matemático

Se inicia este marco teórico con un preámbulo sobre el significado de la palabra modelo y sus implicaciones en el campo de las ciencias exactas. La definición preferida de “modelo” es la siguiente:

“Una representación en miniatura de algo; un patrón de algo por hacer; un ejemplo de imitación o emulación; una descripción o analogía utilizada para ayudar a visualizar algo (por ejemplo, un átomo) que no se puede observar directamente; un sistema de postulados, datos e inferencias presentado como una descripción matemática de una entidad o estado de cosas” (Dym, 2004).

Bender (2012) define que “un modelo matemático es un constructo matemático abstracto, simplificado, relacionado con una parte de la realidad y creado para un propósito particular”.

Los principios del modelado matemático son los siguientes:

- Homogeneidad y consistencia dimensional.
- Abstracción y escala.
- Principios de conservación y equilibrio.
- Construcción de modelos lineales.

### 2.2.2. El modelo estadístico

Para Dobson y Barnett (2018) un buen ajuste debe tomar en cuenta lo siguiente:

- Especificación del modelo: un modelo se especifica en dos partes: una ecuación que vincula la respuesta con las variables explicativas y la distribución de probabilidad de la variable de respuesta.
- Estimación de los parámetros del modelo.
- Verificación de la adecuación del modelo: qué tan bien se ajusta o resume los datos.

- Inferencia: para la inferencia clásica o frecuentista, esto implica calcular intervalos de confianza, probar hipótesis sobre los parámetros en el modelo e interpretar los resultados.

Los autores también sugieren que el análisis de datos debe comenzar con una consideración de cada variable por separado, tanto para verificar la calidad de los datos, como para ayudar con la formulación del modelo.

1. ¿Cuál es la escala de medida? ¿Es continua o categórica? Si es categórica, ¿cuántas categorías tiene?, ¿son nominales u ordinales?
2. ¿Cuál es la forma de la distribución? Esto se puede examinar utilizando tablas de frecuencia, gráficos de puntos, histogramas y otros métodos gráficos.
3. ¿Cómo se asocia una variable con otras? Tabulaciones cruzadas para variables categóricas, diagramas de dispersión para variables continuas, diagramas de caja lado a lado para medidas de escala continua, agrupadas de acuerdo con los niveles del factor de una variable categórica, y otros resúmenes similares pueden ayudar a identificar patrones de asociación. Por ejemplo ¿los puntos en un diagrama de dispersión sugieren asociaciones lineales o no lineales? ¿Las medias de grupo aumentan o disminuyen consistentemente con una variable ordinal, que define los grupos?

## Modelo de regresión lineal simple

Por ejemplo, uno de los modelos más sencillos y frecuentemente utilizados en la literatura, es el modelo de regresión lineal simple. Se puede decir que es el representante más antiguo de la familia del modelo lineal general. Este modelo considera el estudio de una sola variable que explica una sola variable respuesta. La figura 2.1 muestra el diagrama de dispersión de las variables  $x$  e  $y$ , en el entendido de que  $x$  es la variable explicativa (también llamada regresora) e  $y$  es la variable respuesta.

En la figura 2.1, se visualiza un ejemplo sencillo, pero muy utilizado, de modelos matemáticos. Más específicamente, estadísticos. Este modelo se formula una relación entre dos variables. El objetivo es ajustar, mediante mínimos cuadrados ordinarios, la mejor recta, como representación de dicha relación, también

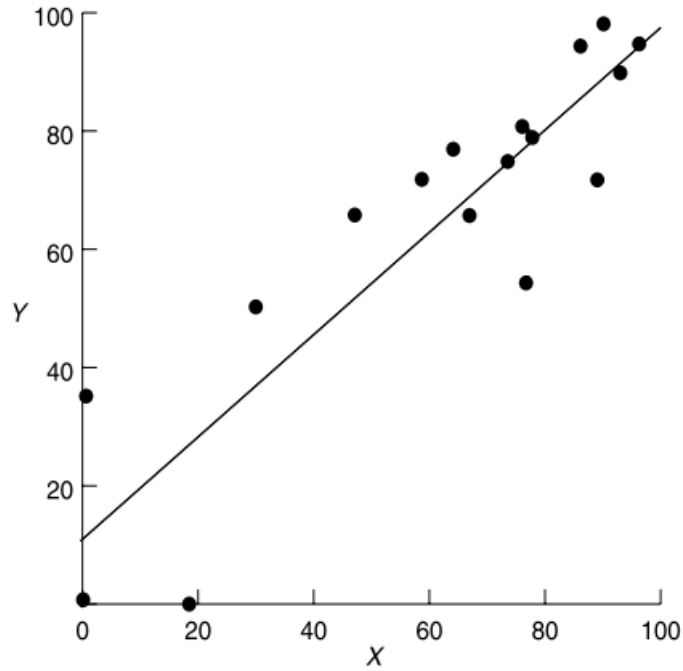


Figura 2.1: Diagrama de dispersión para  $x$  e  $y$  y recta de regresion.  
Fuente: Rencher y Schaalje (2008).

mostrada en la figura. Se plantea:

$$y_i = \alpha + \beta_1 x_i + \epsilon_i \quad (2.1)$$

En (2.1)  $y_i$  se corresponde con la  $i$ -ésima observación de la variable respuesta, para  $i = 1, 2, \dots, n$ , es decir, suponiendo que se cuenta con  $n$  observaciones;  $x_i$  se corresponde con la  $i$ -ésima observación de la variable explicativa ( $x$ );  $\alpha$  se conoce como el intercepto y  $\beta_1$  como el regresor asociado a la variable  $x$ .  $\epsilon_i$  representa el  $i$ -ésimo error no observable, considerado aleatorio, generalmente haciendo la suposición de que se distribuye normalmente con media 0 y varianza  $\sigma^2$ .  $\beta_0, \beta_1$  y  $\sigma^2$  son parámetros cuyos valores se encuentran minimizando la suma de cuadrados del error, a partir de los datos disponibles (Rencher y Schaalje, 2008).

La teoría que sigue sobre el modelo de regresión logística (o logit) es una síntesis del trabajo de Ponsot (2011).

### 2.2.3. La familia exponencial de distribuciones

Sea  $f(y; \theta, \phi)$  una función de densidad, en el caso continuo, o de masa, en el caso discreto, para la variable aleatoria  $Y$ . Se dice que  $f$  pertenece a la familia

exponencial de distribuciones si:

$$f(y; \theta, \phi) = \exp \left[ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right] \quad (2.2)$$

donde  $\theta$  es un parámetro de posición, llamado parámetro canónico, y  $\phi$  es un parámetro de escala. Además,  $a(\cdot)$ ,  $b(\cdot)$  y  $c(\cdot, \cdot)$  son funciones conocidas. Puede demostrarse que  $E[Y] = \mu = b'(\theta)$  y  $V[Y] = \sigma^2 = b''(\theta)a(\phi)$ .

#### 2.2.4. Modelo lineal generalizado (MLG)

Desarrollado originalmente por Nelder y Wedderburn (1972), este modelo supone una equivalencia entre un componente sistemático (lineal como en el modelo lineal general) y una variable respuesta considerada aleatoria, mediante una función de enlace que las relaciona. Este modelo está centrado en la familia exponencial (uniparamétrica) de distribuciones, cuyo parámetro se desconoce y debe ser estimado.

El Modelo Lineal Generalizado (MLG) se define de la siguiente forma:

Sean  $Y_1, Y_2, \dots, Y_n$ ,  $n$  variables aleatorias independientes tales que  $Y_i \sim f(y_i), \forall i$ . Supóngase que cada  $Y_i$  representa la respuesta de un fenómeno en estudio, sobre la que se desea ajustar un modelo estocástico empleando una muestra observada. Sea  $\mu_i = E[Y_i]$  la media correspondiente a la  $i$ -ésima observación de la muestra. Si  $f(\cdot)$  pertenece a la familia exponencial de distribuciones, se denomina **modelo lineal generalizado (MLG)** aquel modelo que, basado en la muestra observada y centrado en el estudio de  $\mu_i$ , postula

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i, \quad i = 1, 2, \dots, n \quad (2.3)$$

donde  $\mathbf{x}_i^T \boldsymbol{\beta} = \eta_i$  se conoce como el componente sistemático del modelo (o predictora lineal) y  $g(\mu_i)$ , transformación uno a uno de la media, continua y diferenciable, se conoce como la función de enlace. Además, en (2.3),  $\mathbf{x}_i^T = [x_{i1} \ x_{i2} \ \dots \ x_{im}]$ , es el  $i$ -ésimo vector fila de la matriz de diseño  $\mathbf{X}_{n \times m}$ , formada por valores no aleatorios seleccionados por el investigador y  $\boldsymbol{\beta} = [\beta_1 \ \beta_2 \ \dots \ \beta_m]^T$  es un vector columna de parámetros desconocidos, cuya estimación se procura a partir de la muestra observada. Luego, un MLG es aquel modelo que asocia un componente aleatorio representado en (2.3) por  $E(Y_i)$  y un componente sistemático ( $\eta_i$ ), a través de una función de enlace  $g(\cdot)$ .

Se trata de una generalización porque la distribución normal (la única considerada en el modelo lineal general) es parte de la familia exponencial, pero en el MLG

también se pueden suponer otras distribuciones como la binomial, la Poisson, la exponencial, la gamma y otras. La realización del MLG que se utiliza en la presente investigación, es la Regresión Logística, también denominada Modelo Logit. Existen otras implementaciones tales como la regresión geométrica, la beta, la gamma, etc. que no serán utilizadas en este estudio.

### 2.2.5. La función de masas binomial

Sea

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, \dots, n \quad (2.4)$$

con  $n$  un entero positivo fijo conocido, y  $p$  un parámetro ( $0 < p < 1$ ). Entonces  $f(\cdot)$ , conocida como la función de masa de probabilidad binomial,  $\text{Bin}(\cdot; n, p)$ , pertenece a la familia exponencial de distribuciones.

Nótese que la binomial no proviene de una variable aleatoria continua, sino discreta, y de hecho pertenece a la familia exponencial de distribuciones. Se pide que  $n$  sea un entero positivo conocido, porque los supuestos operan sobre distribuciones uniparamétricas.

### 2.2.6. El modelo de regresión logística

Se trata de un modelo que permite estimar la probabilidad de éxito (o de fracaso) de la variable respuesta, relacionándola de forma lineal con las características medidas en los individuos. No todos los modelos son consistentes frente a las limitaciones que impone la probabilidad, como respuesta. Al tratarse de una cantidad en el intervalo  $[0, 1]$  [aunque, sin pérdida de generalidad se supondrá  $p \in (0, 1)$ ], el modelo lineal general, por ejemplo, que supone la normalidad de la respuesta definida en  $(-\infty, +\infty)$  no es apropiado. Modelar el logaritmo de la probabilidad, por otro lado, es sin duda mucho más apropiado, pero aún presenta el problema de que cualquier relación funcional que se proponga, debe tener como conjunto de llegada los números reales negativos exclusivamente (o positivos exclusivamente, si se opta por el opuesto), y esto puede dificultar la interpretación de la predictora lineal.

Para subsanar estos inconvenientes es introducida la técnica de regresión logística para variables explicativas de naturaleza continua y su contraparte, el modelo

logit, para variables explicativas categóricas o factores explicativos. Entonces sea

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \quad (2.5)$$

Aplicada a la probabilidad de éxito ( $p$ ) de una variable aleatoria de Bernoulli,  $\text{logit}(p)$  representa el logaritmo de la posibilidad, entendida como el cociente entre la probabilidad de éxito  $p$  y su complemento, la probabilidad de fracaso  $1-p$ . La posibilidad es una medida que toma valores en el intervalo  $[0, +\infty)$ , por consiguiente su logaritmo los toma en el intervalo  $(-\infty, +\infty)$ .

La interpretación de la posibilidad como medida es sencilla. Si el cociente resulta menor que 1, claramente el fracaso tiene mayor posibilidad de ocurrir que el éxito, si por el contrario, resulta mayor que 1, el éxito tiene más posibilidades que el fracaso, y por último, si resulta exactamente igual a 1, el éxito y el fracaso tienen igual posibilidad de ocurrir.

Entonces la formulación del modelo logit o de regresión logística, derivado como un MLG, es la siguiente:

Sean  $Y_1, Y_2, \dots, Y_k$ ,  $k$  variables aleatorias independientes tales que  $Y_i \sim \text{Bin}(y_i; n_i, p_i), \forall i$ . Se conoce como **modelo logit** aquel modelo que postula la equivalencia entre el logaritmo de la posibilidad y una función predictora lineal en los parámetros [como se describe en (2.3)], según la ecuación:

$$\eta_i = \text{logit}(p_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad i = 1, 2, \dots, k \quad (2.6)$$

o equivalentemente, en términos matriciales, como:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \quad (2.7)$$

Como se afirmó previamente, la distribución binomial pertenece a la familia exponencial y, considerando que  $\text{logit}(p_i) = \log[p_i/(1-p_i)] = \log[n_i p_i / (n_i - n_i p_i)] = \log[\mu_i / (n_i - \mu_i)]$  es una función uno a uno, continua y diferenciable de la media  $\mu_i$ ,

el modelo logit es un caso particular del MLG, con las siguientes características:

$$\begin{aligned}
 g(\mu_i) &= \text{logit}(p_i) = \log[\mu_i/(n_i - \mu_i)] = \theta_i \\
 p_i &= \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}} \\
 \mu_i &= n_i p_i \\
 b(\theta_i) &= n_i \log(1 + e^{\theta_i}) \\
 a(\phi) &= 1 \\
 \mathbf{I}(\boldsymbol{\beta}) &= \mathbf{X}^T \mathbf{W} \mathbf{X}
 \end{aligned} \tag{2.8}$$

además,

$$\boldsymbol{\eta} = \begin{bmatrix} \text{logit}(p_1) \\ \text{logit}(p_2) \\ \vdots \\ \text{logit}(p_k) \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_k^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{km} \end{bmatrix}$$

y como antes:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

Para la estimación de los parámetros generalmente se utiliza el **método de mínimos cuadrados iterativos ponderados**. Entonces:

$$\begin{aligned}
 g'(\mu_i) &= \frac{\partial g(\mu_i)}{\partial \mu_i} = \left[ \frac{n_i - \mu_i}{\mu_i} \right] \left[ \frac{(n_i - \mu_i) - \mu_i(-1)}{(n_i - \mu_i)^2} \right] = \frac{n_i}{\mu_i(n_i - \mu_i)} \\
 &= \frac{1}{n_i p_i (1 - p_i)} = (\text{V}[Y_i])^{-1}
 \end{aligned}$$

y de la definición de la familia exponencial de distribuciones:

$$\begin{aligned}
 b'(\theta_i) &= \frac{\partial b(\theta_i)}{\partial \theta_i} = \frac{ne^{\theta_i}}{1 + e^{\theta_i}} = n_i p_i = \text{E}[Y_i] \\
 b''(\theta_i) &= \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} = \frac{ne^{\theta_i}}{[1 + e^{\theta_i}]^2} = n_i p_i (1 - p_i) = \text{V}[Y_i]
 \end{aligned}$$



Luego  $w_i = \{b''(\theta_i)[g'(\mu_i)]^2\}^{-1} = \{V[Y_i](V[Y_i])^{-2}\}^{-1} = V[Y_i]$ , y

$$\mathbf{W} = \begin{bmatrix} V[Y_1] & 0 & \cdots & 0 \\ 0 & V[Y_2] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V[Y_k] \end{bmatrix} \quad \mathbf{\Delta} = \begin{bmatrix} (V[Y_1])^{-1} & 0 & \cdots & 0 \\ 0 & (V[Y_2])^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & (V[Y_k])^{-1} \end{bmatrix}$$

Entonces  $\mathbf{W}\mathbf{\Delta} = \mathbf{I}_{(k)}$  (la matriz identidad de dimensión  $k$ ). Finalmente, la estimación máximo-verosímil de los parámetros se obtiene iterando sobre la ecuación que sigue:

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \boldsymbol{\mu}) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(r)}} \quad (2.9)$$

Se parte de la condición inicial  $\boldsymbol{\beta}^{(0)} : g(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}^{(0)}$ , esto es, haciendo el vector de parámetros inicial como la solución del sistema de ecuaciones derivado del modelo, en el que la media se aproxima por los valores observados de la variable aleatoria.

La principal característica del modelo logit es la posibilidad que brinda de estudiar los efectos que tienen los factores sobre la respuesta, además del sentido de dicho efecto.

### 2.2.7. Pruebas de hipótesis e intervalos de confianza sobre $\hat{\boldsymbol{\beta}}$

Las pruebas de hipótesis que se realizan sobre los parámetros en un MLG, son esencialmente de dos tipos: pruebas de Wald y pruebas de la razón de verosimilitud. Ambas conducen a la misma distribución asintótica  $\chi^2$  para efectos de la decisión.

La **prueba de Wald** se apoya en la propiedad de normalidad asintótica de los estimadores máximo-verosímiles, aprovechando las varianzas estimadas en la diagonal de la matriz de varianzas y covarianzas de  $\hat{\boldsymbol{\beta}}$  (llámense  $\hat{\sigma}_{ii}^2, i = 1, \dots, m$ ). En efecto, para cualquier  $i$ , sea la hipótesis nula  $H_0 : \beta_i = \beta_0$ , en la cual  $\beta_i$  es uno de los parámetros de interés y  $\beta_0$  es una constante. El estadístico

$$Wd = \frac{(\hat{\beta}_i - \beta_0)^2}{\hat{\sigma}_{ii}^2} \sim \chi_1^2 \quad (2.10)$$

Consecuentemente, siguiendo la prueba de Wald, se rechaza  $H_0$  si  $Wd > \chi_{1,1-\alpha}^2$ , donde  $\alpha$  es el nivel de significación escogido. Obviamente, un intervalo de Wald del  $(1 - \alpha)\%$  de confianza para  $\beta_i$  es el conjunto de todos los  $\beta_0 : Wd \leq \chi_{1,1-\alpha}^2$ , o equivalentemente,  $\beta_i = \hat{\beta}_i \pm \nu \hat{\sigma}_{ii}$  donde, como se acostumbra,  $\nu$  es la abscisa que acumula una probabilidad  $1 - \alpha/2$  en la distribución normal estándar.

Por su parte, la **prueba de la razón de verosimilitud** forma el cociente entre el máximo de la función de verosimilitud bajo la hipótesis nula y el máximo de la función de verosimilitud irrestricta. Sea la hipótesis nula, como antes,  $H_0 : \beta_i = \beta_0$  para algún  $i$ . El máximo de la verosimilitud restringida a la hipótesis nula es  $L(\tilde{\beta}_1, \dots, \tilde{\beta}_i = \beta_0, \dots, \tilde{\beta}_m)$ , mientras que el máximo de la verosimilitud irrestricta es  $L(\hat{\beta}_1, \dots, \hat{\beta}_m)$ . La razón de verosimilitudes es  $\Lambda = L(\tilde{\beta}_1, \dots, \tilde{\beta}_i = \beta_0, \dots, \tilde{\beta}_m) / L(\hat{\beta}_1, \dots, \hat{\beta}_m)$ . Por definición,  $L(\tilde{\beta}_1, \dots, \tilde{\beta}_i = \beta_0, \dots, \tilde{\beta}_m) \leq L(\hat{\beta}_1, \dots, \hat{\beta}_m)$ , entonces  $\Lambda \leq 1$  y, asintóticamente, el estadístico

$$\begin{aligned} \lambda &= -2 \log \Lambda \\ &= -2[l(\tilde{\beta}_1, \dots, \tilde{\beta}_i = \beta_0, \dots, \tilde{\beta}_m) - l(\hat{\beta}_1, \dots, \hat{\beta}_m)] \sim \chi_1^2 \end{aligned} \quad (2.11)$$

Consecuentemente, siguiendo la prueba de la razón de verosimilitud, se rechaza  $H_0$  si  $\lambda > \chi_{1,1-\alpha}^2$ , donde  $\alpha$  es nuevamente el nivel de significación escogido. Un intervalo del  $(1 - \alpha)\%$  de confianza según la razón de verosimilitud para  $\beta_i$  es el conjunto de todos los  $\beta_0 : \lambda \leq \chi_{1,1-\alpha}^2$ .

## 2.2.8. Pruebas de bondad del ajuste

En todo proceso de modelado estadístico es necesario contar con una forma de medir qué tan bien ajusta el modelo propuesto los datos disponibles. Este tipo de pruebas de hipótesis se conocen como pruebas de bondad del ajuste. Para el MLG las pruebas de bondad del ajuste más extendidas son:

- La prueba  $\chi^2$  de Pearson.
- El análisis del deviance.

La prueba  **$\chi^2$  de Pearson** construye el estadístico (Hosmer y Lemeshow, 2000, p. 145):

$$r(y_1, \dots, y_k; \hat{y}_1, \dots, \hat{y}_k) = \sum_{i=1}^k \frac{(y_i - \hat{y}_i)^2}{\hat{V}[Y_i]} \quad (2.12)$$

que se distribuye asintóticamente  $\chi^2$  con  $k - m$  grados de libertad. En (2.12)  $\hat{y}_i$  simboliza el  $i$ -ésimo valor predicho por el modelo para la correspondiente variable aleatoria  $Y_i$  ( $i = 1, \dots, k$ ), por lo tanto, la cantidad  $y_i - \hat{y}_i$  es el  $i$ -ésimo residuo observado. Luego, la ecuación (2.12) es la suma del cuadrado de los residuos estandarizados de Pearson, de donde obtiene su nombre la prueba de hipótesis en cuestión. La prueba de bondad del ajuste del modelo propuesto, para un nivel de significación  $\alpha$ , consiste en rechazar su idoneidad si  $r(\cdot; \cdot) > \chi_{k-m, 1-\alpha}^2$ , y aceptarla en caso contrario.

Por su parte, se conoce como *deviance* ( $D$ ) al estadístico formado a partir de la prueba del logaritmo de la razón de verosimilitudes máximas, en la comparación de un modelo en observación (llámese  $M_o$ ) con  $m_o$  parámetros, versus el modelo saturado (llámese  $M_s$ ) con  $k$  parámetros ( $m_o < k$ ). Esto es:

$$D = -2[l_{M_o}(\tilde{\beta}_1, \dots, \tilde{\beta}_{m_o}) - l_{M_s}(\hat{\beta}_1, \dots, \hat{\beta}_k)] \quad (2.13)$$

En (2.13),  $D$  se distribuye asintóticamente  $\chi^2$  con  $k - m_o$  grados de libertad. El número de parámetros estimados bajo  $M_o$  es menor que el número de parámetros estimados bajo  $M_s$ , por lo tanto, la prueba de hipótesis con base en el *deviance* es una prueba de la nulidad de aquellos parámetros presentes en  $M_s$  y ausentes en  $M_o$ . Consecuentemente, si el modelo en observación ( $M_o$ ) ajusta apropiadamente los datos,  $D$  será un valor razonablemente pequeño, luego, la prueba de bondad del ajuste del modelo en observación mediante el *deviance*, con un nivel de significación  $\alpha$ , consiste en rechazar su idoneidad si  $D > \chi_{k-m_o, 1-\alpha}^2$ , y aceptarla en caso contrario.

Por otra parte, aún cuando el *deviance* es una prueba que contrasta un modelo en observación versus el modelo saturado, éste último, con ajuste perfecto de los parámetros, el procedimiento también puede emplearse para contrastar cualesquiera dos modelos que se postulan sobre la situación de datos. En general, esta propiedad se aprovecha para contrastar modelos que incorporan uno a uno parámetros hasta alcanzar el modelo saturado. De esta forma se cuenta con un procedimiento que apoya al investigador en la decisión del mejor modelo posible apropiado para el problema que enfrenta. Sean  $M_s$  el modelo saturado,  $M_a$  con  $a$  parámetros y  $M_b$  con  $b$  parámetros, modelos alternativos tales que los parámetros de  $M_a$  están contenidos en  $M_b$  y  $a < b \leq s$ . Entonces,

$$\begin{aligned} D_{ab} &= D_a - D_b = -2[l_{M_a}(\cdot) - l_{M_s}(\cdot)] - (-2[l_{M_b}(\cdot) - l_{M_s}(\cdot)]) \\ &= -2[l_{M_a}(\cdot) - l_{M_b}(\cdot)] \end{aligned} \quad (2.14)$$

Así, para comparar los modelos alternativos  $M_a$  y  $M_b$  puede emplearse el análisis del *deviance*, que para el caso se distribuye según una  $\chi^2$  con  $b - a$  grados de libertad. La hipótesis a contrastar es que  $M_a$  ajusta tan bien los datos como los ajusta  $M_b$ , y se rechaza al nivel de significación  $\alpha$  si  $D_{ab} > \chi_{b-a, 1-\alpha}^2$ .

### 2.2.9. Técnicas de selección de variables

En el modelo logit, con el fin de orientar el proceso de selección de variables explicativas, generalmente se utiliza uno de los tres métodos siguientes:

- **Forward:** Selección de variables hacia adelante, se trata de una eliminación progresiva. Se inicia con el modelo más sencillo, luego se van añadiendo variables que son aceptadas con ciertos criterios, hasta que ya no se encuentran variables que los cumplan. El modelo sugerido es el que contiene las variables que lograron ser incluidas.
- **Backward:** Elimina variables hacia atrás, se trata de una eliminación regresiva, lo contrario al Forward. Se inicia desde el modelo saturado (aquel que contiene todas las variables posibles) y en cada paso se elimina la variable que menos efectos tiene sobre la respuesta, atendiendo también a ciertos criterios. El modelo sugerido es el que contiene las variables que no lograron ser excluidas.
- **Stepwise:** Método por pasos que consiste en la selección automática de variables más significativas, inclusión / exclusión de manera secuencial utilizando sólo el criterio de la significación estadística.

Los criterios señalados generalmente están basados en el **AIC** (por las siglas en inglés de *Akaike Information Criteria*). Este criterio se basa en la ya mencionada verosimilitud, pero penaliza el número de parámetros utilizados en los modelos bajo estudio. Aquel modelo que tenga el menor AIC es el mejor modelo.

### 2.2.10. Curvas ROC

Una curva ROC (por las siglas en inglés de *Receiver Operating Characteristic*) es una técnica gráfica muy extendida, para evaluar la capacidad de clasificación de un modelo con respuesta dicotómica (Fawcett, 2006). Supóngase que a partir de un modelo clasificador se cuenta con la matriz de confusión de la tabla 2.1.

Tabla 2.1: Matriz de confusión

El clasificador	La realidad		
	Positivo	Negativo	Total
Positivo	VP	FP	TPC
Negativo	FN	VN	TNC
Total	TPR	TNR	T

Fuente: Fawcett (2006).

En la tabla 2.1, V abrevia “Verdaderos”, F abrevia “Falsos”, P “Positivos”, N “Negativos”, T “Total”, R “Filas” (por las siglas en inglés de *Row*) y C “Columnas”. Así, por ejemplo, VN significa “Verdaderos Positivos” y se refiere al número de clasificaciones detectadas por el modelo como negativas, que efectivamente lo fueron.

La curva ROC se construye graficando, a partir de la tabla 2.1, la tasa de verdaderos positivos (TsVP) -también llamada sensibilidad- versus la tasa de falsos positivos (TsFP). Ahora “Ts” abrevia el vocablo “Tasa”, entonces:

$$TsVP = \frac{VP}{TPR} = \frac{VP}{VP + FN}$$

$$TsFP = \frac{FP}{TNR} = \frac{FP}{FP + VN}$$

Estas expresiones con frecuencia se encuentran en términos de la la tasa de verdaderos negativos (TsVN), también conocida como especificidad, dada como:

$$TsVN = \frac{VN}{TNR} = \frac{VN}{FP + VN} = 1 - TsFP$$

Luego, la curva ROC se construye graficando la sensibilidad versus uno menos la especificidad. El objetivo mostrar si el clasificador produce más resultados “buenos”, que “malos”. Cualquier clasificador dicotómico produce un punto en este espacio bidimensional, por lo tanto se trata de estimar varios puntos. De manera que para producir una curva, se requiere alterar sistemáticamente la condición de clasificación, Por ejemplo, si la clasificación se produce mediante una probabilidad  $p$ , evaluando  $p > p_0$  o  $p < p_0$ , la idea es variar  $p \in (0, 1)$ .

El área bajo la curva ROC se utiliza para comparar entre clasificadores. El valor 1 corresponde a un clasificador perfecto (todos los resultados fueron correctos), el 1/2 a un clasificador aleatorio y el valor 0 a uno totalmente imperfecto. Este último concepto en realidad no existe, puesto que si un clasificador se desempeña

totalmente mal, basta con invertir el resultado de la clasificación, para tornarlo en uno que se desempeña de forma perfecta.

### 2.2.11. Emprendimientos

Es importante en la presente investigación, la aplicación de los modelos de regresión logística a los emprendimientos productivos, por lo que se definen los factores del éxito empresarial (ver 2.2.

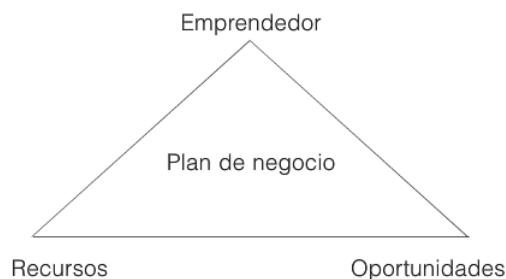


Figura 2.2: Factores del éxito empresarial  
Fuente: Vainrub (2009).

Los tres factores son fundamentales, sin embargo, cuando un financista experimentado estudia una propuesta de negocios, le otorga generalmente mayor peso al emprendedor. La razón es que el mundo es dinámico, las ideas cambian, las oportunidades desaparecen y los recursos pueden ser volátiles. Los buenos empresarios se adaptan a los cambios, modifican sus ideas y buscan nuevas oportunidades o nuevos recursos. Es más difícil cambiar al individuo que a las ideas o los recursos. Por ello el emprendedor es el elemento más importante de la tríada, sin olvidar que debe apoyarse en una sólida base de recursos y oportunidades, y guiarse por un plan coherente Vainrub (2009).

Según Alcaraz (2011) el emprendedor posee características que le facilitan no sólo emprender, sino también hacerlo con éxito. Entre ellas, algunas de las más mencionadas son: creatividad e innovación, confianza en él mismo y sus capacidades, perseverancia, capacidad para manejar problemas y aceptación del riesgo.

Como muestra la figura 2.3, en 2019 Ecuador presentó una Actividad Emprendedora Temprana (TEA) de 36.2%, superior a la de 2017 que fue de 29.62%. Esto significa que, en Ecuador, alrededor de 3.6 millones de habitantes estuvieron involucrados en la puesta en marcha de un negocio. El aumento de la TEA se visualiza mejor a través de sus componentes: la tasa de emprendedores nacientes en 2019 fue 26.9% y la de negocios nuevos, 10.8%; mientras que estas tasas en

2017 fueron 21.2 y 9.8, respectivamente. Es decir, el cambio en 2019 se da principalmente porque hay más personas tomando acciones para poner en marcha un negocio que aún no está en funcionamiento Lasio, Amaya, Zambrano, y Ordeñana (2020).

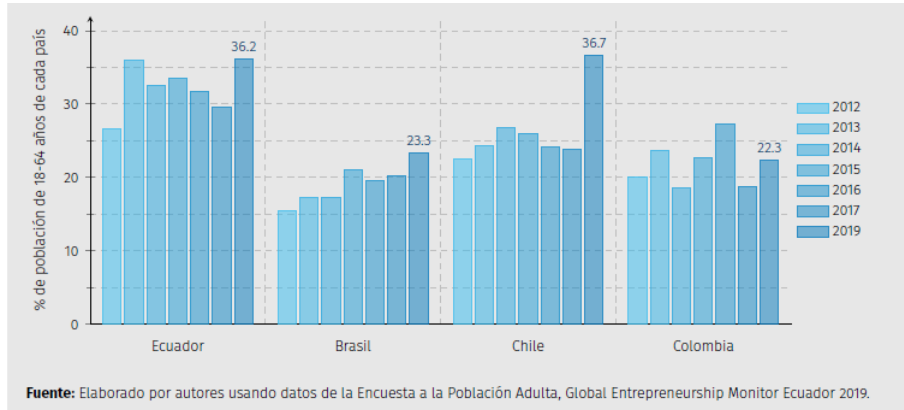


Figura 2.3: Evolución de la Actividad Emprendedora Temprana en Ecuador y países seleccionados  
Fuente: Lasio y cols. (2020).

# Capítulo 3

## Marco metodológico

### 3.1. Ubicación

La presente investigación se basa en los datos del estudio de sostenibilidad de los emprendimientos productivos de la Provincia de Tungurahua Frente 1 - Fase 2, desarrollado por la Facultad de Contabilidad y Auditoría de la Universidad Técnica de Ambato, mediante la Dirección de Vinculación con la Sociedad. El autor fue investigador principal y partícipe directo del proyecto mencionado. La Fase 2 se relaciona con cinco cantones de la Provincia del Tungurahua: Cevallos, Quero, Mocha, Patate y Tisaleo.

### 3.2. Equipos y materiales

Para la preparación de los datos, se utilizó un equipo de cómputo portátil robusto con Microsoft Excel. Durante la investigación se dispuso de la base general actualizada de la Provincia de Tungurahua y las bases de datos independientes por cada cantón y por parroquia. Por otra parte, la codificación se ha desarrollado usando software libre, más específicamente utilizando **R** (R Core Team, 2020), que es una herramienta muy robusta para Estadística y que no necesita del pago de licencia. El código fuente programado se presenta en el Anexo 2.



### **3.3. Tipo de investigación**

La presente investigación se centra en el paradigma crítico, por su análisis crítico en el proceso investigativo, basado en la acción práctica y válida en el tiempo, así como propositivo porque plantea alternativas de solución factible al problema. Además, tiene un enfoque cuantitativo, según Hernández, Fernández, y Baptista (2014), ya que parte de una idea que va acotándose y, una vez delimitada, se derivan objetivos y preguntas de investigación. Se revisa la literatura y se construye un marco o una perspectiva teórica; de las preguntas se establecen hipótesis y determinan variables; se traza un plan para probarlas (diseño); se miden las variables en un determinado contexto; se analizan las mediciones obtenidas utilizando métodos estadísticos y se extraen conclusiones.

Además, por el uso de la matemática y la estadística se obtiene la confiabilidad en la medición numérica, para establecer con exactitud ciertos comportamientos en base a diversas características, como variables explicativas.

De la misma manera la presente investigación al ser del tipo cuantitativo, dispone de un alcance exploratorio, correlacional, descriptivo y explicativo (Hernández y cols., 2014).

Se pretende validar el conocimiento en base a las hipótesis formuladas, características causales y sus efectos, dando así alternativas de solución en la toma de decisiones, para el beneficio de los emprendedores en sus emprendimientos productivos.

### **3.4. Pruebas de Hipótesis**

#### **3.4.1. Preguntas que motivan la investigación**

- ¿Es o no posible explicar el éxito y el fracaso de un emprendimiento productivo de la provincia de Tungurahua, utilizando las variables: actividad, instrucción, edad, género, tiempo del emprendimiento, aspectos, problemas, publicidad, personal, cantón, si ha sido suspendido, o cerrado?
- ¿Es factible ajustar un modelo matemático/estadístico que permita estudiar este fenómeno?
- ¿En qué sentido y magnitud las variables consideradas explicativas afectan el resultado de un emprendimiento productivo?

### 3.4.2. Sistema de Hipótesis

- Existen factores que explican el éxito de los emprendimientos productivos en los cantones de la Provincia de Tungurahua.
- La zona geográfica cantón en donde se desarrolla su actividad, actividad, instrucción, edad, género, tiempo emprendimiento, aspectos, problemas, publicidad, personal, suspendido, cerrado, contribuyen a explicar el éxito o el fracaso de los emprendimientos productivos.
- El presupuesto promedio incide de manera significativa en el éxito de los emprendimientos productivos.
- El nivel de estudio incide en el éxito de los emprendimientos productivos.
- El ciclo de vida del emprendedor (tiempo que lleva en su emprendimiento) influye significativamente en el éxito.
- Los emprendimientos suspendidos temporalmente explican el fracaso de los emprendimientos.
- El personal con el que cuenta bajo relación de dependencia explica el fracaso de los emprendimientos productivos.
- La publicidad, influye en el éxito de los emprendimientos.

### 3.5. Población y muestra

La presente investigación utiliza los datos obtenidos del proyecto académico de Prácticas Pre-profesionales de Servicio a la Comunidad de Vinculación con la sociedad No. 3.9, denominado: “La sostenibilidad de los emprendimientos productivos de la provincia de Tungurahua, frente 2, fase I”. El autor fue docente participante - tutor de dicho proyecto.

Se utilizó un muestreo por conglomerados, agrupamientos naturales relativamente homogéneos en la población. Se realizaron muestras independientes de los cinco cantones: Cevallos, Quero, Mocha, Patate y Tisaleo, según se detalla en la tabla 3.1. Dichos datos utilizados se basan en la información pública de la página del SRI con fecha a diciembre 2019.

Del total de las muestras, se procedió a realizar la preparación y limpieza de datos por cantón y resultó un conjunto formado solamente con las encuestas que fueron

Tabla 3.1: Población y muestra por cantones

Cantón	Población	Muestra
Cevallos	651	353
Quero	1661	520
Mocha	537	225
Patate	1447	494
Tisaleo	1207	501
Total		2093

Fuente: Elaboración propia.

completamente contestadas por los emprendedores, al menos para las variables de interés. En total **1546** emprendimientos productivos.

### 3.5.1. Recolección de información

El instrumento aplicado fue desarrollado, revisado y aprobado entre los entes cooperantes: Consejo Nacional de Competencias; Honorable Gobierno Provincial De Tungurahua; CONAGOPARE, BANECUADOR, Cámara de Comercio de Ambato (ver Anexo 1).

## 3.6. Procesamiento de la información y análisis estadístico

En el estudio original existen alrededor de 24 variables (ver Anexo 1), pero para el presente trabajo se fueron eliminando variables como algunas opiniones y expectativas, que no tienen mayor incidencia en los objetivos.

El modelo estadístico aplicado es la Regresión Logística, basado en el Modelo lineal Generalizado. Este modelo involucra una variable dependiente  $y$  (“cerrado”), dicotómica, es decir, que toma sólo dos valores: 1 si se trata de un éxito y 0 si se trata de un fracaso. Para el presente caso, se trata de la situación del negocio, si ha cerrado definitivamente (éxito) o no (fracaso). Las variables independientes, llamadas también explicativas, son numéricas y categóricas como podemos observar en la tabla 3.2. Se dispone de 12 variables con las que se busca explicar el éxito o fracaso de los emprendimientos productivos.

Tabla 3.2: Variables y tipos

Variable	Clase	Niveles
Presupuesto	factor	0 - 300, 301 - 600, 601 - 900, Más de 900
Actividad	factor	Actividades de apoyo a la agricultura y ganadería y actividades, Cría de aves de corral, Cría de ganado y bovino , Cultivo de plantas no perennes, Cultivo de plantas perennes, Elaboración de alimentos preparados, Elaboración de bebidas, Elaboración de especias salsas, Elaboración de otros productos , Elaboración de productos lácteos, Elaboración y Conservación , Fabricación de calzado, Fabricación de diversos productos de madera , Fabricación de prendas de vestir, Fabricación y comercialización de artículos de construcción, Fabricación y comercialización de otros productos , Ganadería, Otro actividad en general, Producción de carne y de productos carnicos , Tejedura de Productos Textiles, Tropicales y Subtropicales, Venta al por mayor de materiales para la construcción, artículos de ferretería, equipo, materiales de fontanería o gasfitería, Venta al por mayor de materiales para la construcción, artículos de ferretería, equipo, materiales de fontanería o gasfitería
Instrucción	factor	Primaria, Secundaria, Sin Instrucción, Tercer Nivel
Edad	integer	NA
Género	factor	Femenino, Masculino
Tiempo _emprende	factor	Hasta 1 año, Hasta 2 años, Hasta 3 años, Hasta 4 años en adelante
Aspectos _emprende	factor	Acceso a financiamiento, El tiempo para obtener ingresos o ganancias, La experiencia y conocimientos, La facilidad de producción, Los riesgos que conllevan
Problemas	factor	Competencia, Costo de mano de obra, Costo de materia prima, Costo de transporte, Ninguna, Otra, Problemas familiares
Publicidad	factor	Ninguno, Otros, Páginas web, Prensa escrita, Radio, Redes sociales
Suspendido	factor	No, Si
Cerrado	factor	No, Si
Personal	factor	1 persona, 11 a 20 personas, 21 personas o mas, 5 a 10 personas
Cantón	factor	CEVALLOS, MOCHA, PATATE, QUERO, TISALEO

Fuente: Elaboración propia.

# Capítulo 4

## Resultados y discusión

### 4.1. Análisis Descriptivo de los datos

#### 4.1.1. Análisis unidimensional

##### Género

Hay más emprendimientos productivos dirigidos por hombres que por mujeres, como es posible apreciar en la tabla 4.1. En la figura 4.1 se ilustran los porcentajes. Esto puede deberse a que, como se verá más adelante, la mayoría sus actividades se relacionan con la ganadería y la fabricación de calzado, entre otras que requieren de esfuerzo físico. En esta y las tablas que siguen, f abrevia “frecuencia absoluta”, F abrevia “frecuencia absoluta acumulada”, rf abrevia “frecuencia relativa” y rF abrevia “frecuencia relativa acumulada.”

Tabla 4.1: Distribución de frecuencias para la variable Género

Género	f	F	rf	rF
Masculino	959	959.00	0.62	0.62
Femenino	587	1546.00	0.38	1.00

Fuente: Elaboración propia.

##### Edad

Observamos en la tabla 4.2 que la clase modal de la edad de los emprendedores es [44,5, 50,8). De hecho su edad promedio es de aproximadamente 47 años. La

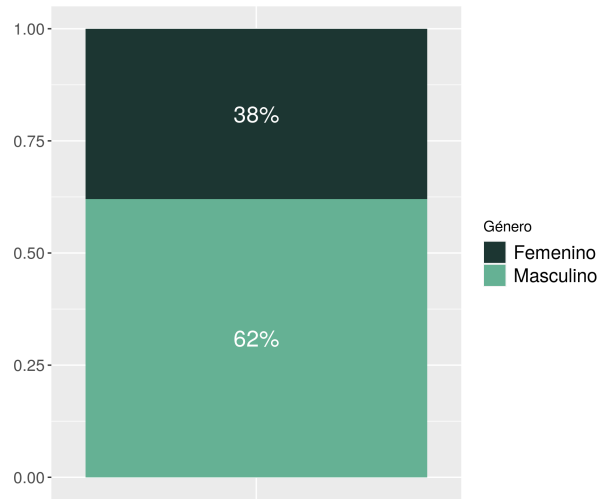


Figura 4.1: Distribución por género.  
Fuente: Elaboración propia.

Tabla 4.2: Distribución de frecuencias para la variable Edad

Edad	f	F	rf	rF
[19,25.4)	23	23.00	0.01	0.01
[25.4,31.7)	73	96.00	0.05	0.06
[31.7,38.1)	244	340.00	0.16	0.22
[38.1,44.5)	264	604.00	0.17	0.39
[44.5,50.8)	403	1007.00	0.26	0.65
[50.8,57.2)	315	1322.00	0.20	0.86
[57.2,63.5)	129	1451.00	0.08	0.94
[63.5,69.9)	58	1509.00	0.04	0.98
[69.9,76.3)	24	1533.00	0.02	0.99
[76.3,89)	12	1545.00	0.01	1.00
[89,Inf)	1	1546.00	0.00	1.00

Fuente: Elaboración propia.

mayor concentración de edades entre los entrevistados oscila entre 40 y 60 años, como se aprecia en la figura 4.2.

### Nivel de instrucción

Es importante mencionar que el nivel de instrucción es un pilar fundamental para el éxito de un emprendimiento, sin embargo, como se ve en la figura 4.3, la mayoría tienen un nivel muy bajo de estudios (primaria y secundaria). Quienes tienen estudios superiores apenas llegan al 6% de la muestra, como puede apreciarse en la tabla 4.3. Esto puede deberse a que, desde hace muchos años atrás, los emprendedores que trabajan en el campo en su mayoría no tienen acceso a la educación superior. Sin embargo, la experiencia ha fortalecido sus negocios.

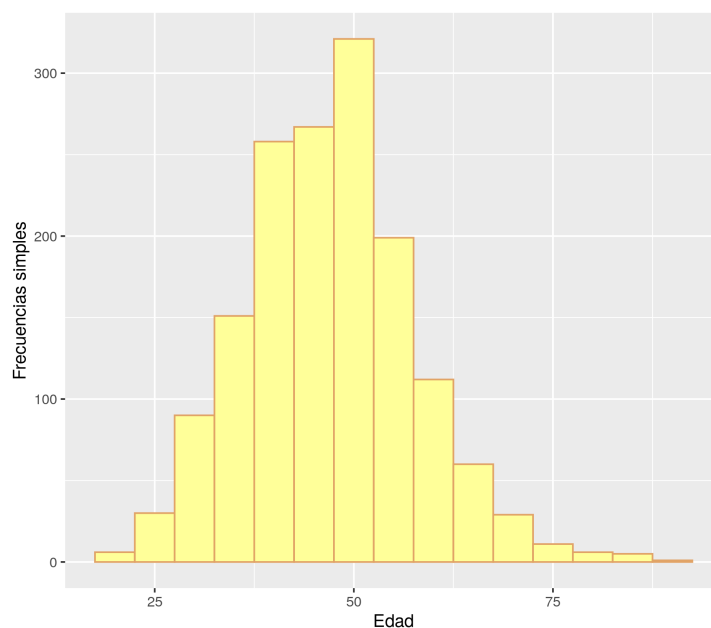


Figura 4.2: Distribución por edad.

Fuente: Elaboración propia.

Tabla 4.3: Distribución de frecuencias para la variable Instrucción

Instrucción	f	F	rf	rF
Primaria	755	755.00	0.49	0.49
Secundaria	678	1433.00	0.44	0.93
Tercer Nivel	97	1530.00	0.06	0.99
Sin Instrucción	16	1546.00	0.01	1.00

Fuente: Elaboración propia.

## Presupuesto

Tabla 4.4: Distribución de frecuencias para la variable Presupuesto

Presupuesto	f	F	rf	rF
Más de 900	616	616.00	0.40	0.40
301 - 600	359	975.00	0.23	0.63
601 - 900	345	1320.00	0.22	0.85
0 - 300	226	1546.00	0.15	1.00

Fuente: Elaboración propia.

Una de las características importantes que pueden explicar el modelo es el presupuesto promedio anual, es decir el capital invertido de los emprendedores para sus actividades (ver tabla 4.4 y figura 4.4). Invierten más de 900 dólares como promedio anual, lo que les ayuda a que las ideas de negocios se puedan generar, pero es evidente que se necesita el apoyo del gobierno y de la banca, mediante sus entidades cooperantes.

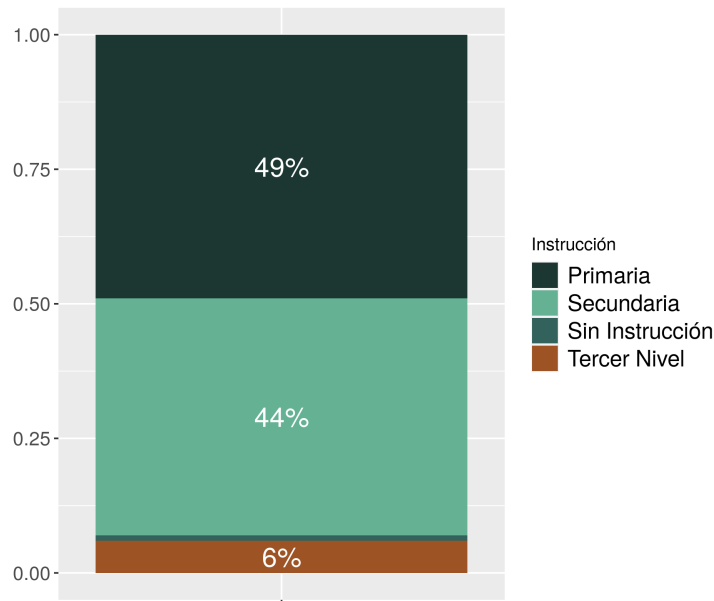


Figura 4.3: Nivel de instrucción de los emprendedores.

Fuente: Elaboración propia.

El presupuesto generalmente se compone de créditos que obtienen los productores. Algunas entidades cooperantes, están trabajando para que se pueda adelantar un año, como parte de varios proyectos que están planteados.

Por otro lado, se detectan emprendimientos, seguramente de las personas de bajos recursos económicos, cuya inversión está por debajo de un salario mínimo básico.

## Tiempo

Tabla 4.5: Distribución de frecuencias para la variable Tiempo\_emprende

Tiempo_emprende	f	F	rf	rF
Hasta 4 años en adelante	880	880.00	0.57	0.57
Hasta 3 años	288	1168.00	0.19	0.76
Hasta 2 años	255	1423.00	0.16	0.92
Hasta 1 año	123	1546.00	0.08	1.00

Fuente: Elaboración propia.

Es fácil identificar que la mayoría de los emprendimientos tienen una trayectoria de muchos años. Al revisar el modelo se verá si este factor tiende al éxito, considerando la experiencia de cada emprendedor con sus actividades. Como se muestra en la tabla 4.5 y la figura 4.5, se nota claramente que existen nuevos negocios que están iniciando su actividad, pero la mayoría de los emprendimientos productivos están entre cuatro años o más.



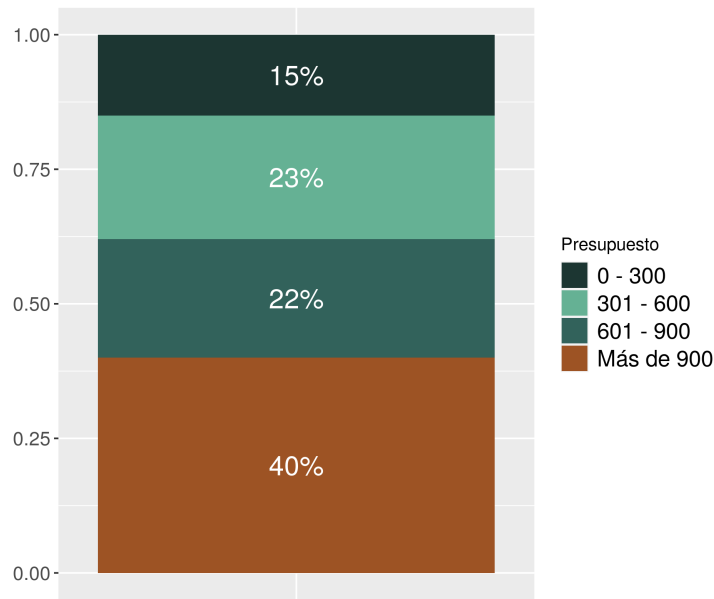


Figura 4.4: Presupuesto promedio anual de los emprendedores.  
Fuente: Elaboración propia.

### Actividades

La tabla 4.6 muestra las distintas actividades a las que se dedican los entrevistados. Allí puede observarse que aproximadamente el 60 % se dedican a actividades agropecuarias. La zona en estudio tiene entonces una fuerte vocación al trabajo del campo.

### Aspectos para el emprendimiento

Con respecto a los aspectos considerados para el emprendimiento, la mayoría se inclina por la experiencia y los conocimientos, como de vital importancia para explicar si esto conlleva al éxito de su negocio. Como se observa en la tabla 4.7 y la figura 4.6, cerca del 51 % indicaron estos aspectos como los más importantes. Sorprende que los entrevistados no consideran el financiamiento un aspecto muy importante, de hecho, en su opinión es el de menor importancia (lo mencionan cerca del 6 % apenas).

### Problemas en el emprendimiento

En cuanto a los problemas para el desarrollo del emprendimiento, es notable que la competencia es el factor más importante para los entrevistados (tabla 4.8 y figura

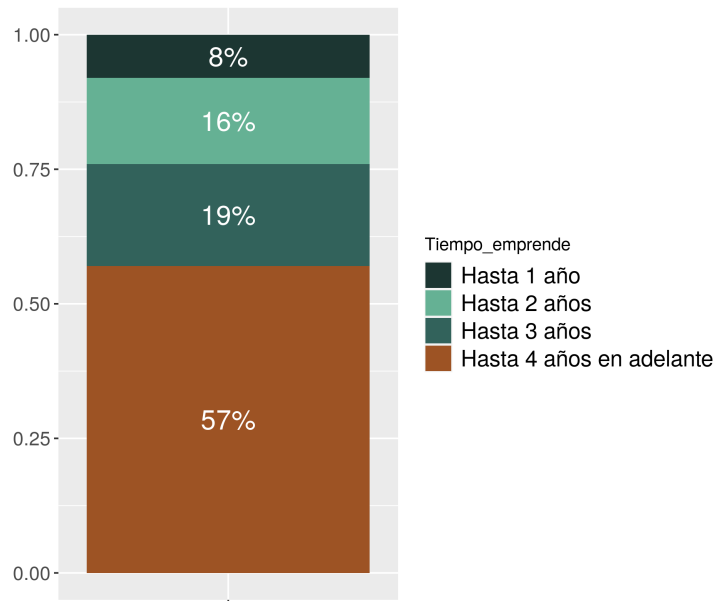


Figura 4.5: Tiempo que lleva el emprendimiento.

Fuente: Elaboración propia.

4.7). Debe considerarse que la mayoría de las actividades son similares dentro sus sectores. Además, el costo de la materia prima sería otro de los problemas que mencionan, pero la necesitan para elaborar sus productos.

## Publicidad

En relación con el tipo de publicidad como se promocionan los emprendimientos productivos, se puede observar que tanto las redes sociales y la difusión mediante radio son los promotores intermediarios en la venta de los productos (tabla 4.9 y figura 4.8). El uso de la tecnología hoy en día es una fuente de publicidad que no necesita de mucha inversión. Un ejemplo de las redes sociales, es la época actual de la emergencia sanitaria, en la que han tenido cabida diferentes negocios que jamás se hubieran dado a conocer.

La publicidad mediante radio, corresponde a radios del mismo sector, en las que sí se necesita invertir. Además, se puede observar que la opción otros puede ser la publicidad denominada de boca en boca, en la cual los mismos clientes o familiares tratan de comunicar la existencia del emprendimiento.

Tabla 4.6: Distribución de frecuencias para la variable Actividad

Actividad	f	F	rf	rF
Cultivo de plantas no perennes	345	345.00	0.22	0.22
Cultivo de plantas perennes	232	577.00	0.15	0.37
Cría de ganado y bovino	189	766.00	0.12	0.50
Fabricación de calzado	141	907.00	0.09	0.59
Otro actividad en general	141	1048.00	0.09	0.68
Fabricación de prendas de vestir	101	1149.00	0.07	0.74
Ganadería	80	1229.00	0.05	0.79
Fabricación y comercialización de otros productos	73	1302.00	0.05	0.84
Venta al por mayor de materiales para la construcción, artículos de ferretería, equipo, materiales de fontanería o gasfitería	51	1353.00	0.03	0.88
Fabricación de diversos productos de madera	41	1394.00	0.03	0.90
Elaboración de productos lácteos	32	1426.00	0.02	0.92
Elaboración de otros productos	28	1454.00	0.02	0.94
Actividades de apoyo a la agricultura y ganadería y actividades	24	1478.00	0.02	0.96
Cría de aves de corral	22	1500.00	0.01	0.97
Elaboración de alimentos preparados	16	1516.00	0.01	0.98
Tejedura de Productos Textiles	11	1527.00	0.01	0.99
Elaboración de bebidas	7	1534.00	0.00	0.99
Producción de carne y de productos carnicos	5	1539.00	0.00	1.00
Elaboración y Conservación	4	1543.00	0.00	1.00
Elaboración de especias salsas	1	1544.00	0.00	1.00
Fabricación y comercialización de artículos de construcción	1	1545.00	0.00	1.00
Tropicales y Subtropicales	1	1546.00	0.00	1.00

Fuente: Elaboración propia.

Tabla 4.7: Distribución de frecuencias para la variable Aspectos\_emprende

Aspectos_emprende	f	F	rf	rF
La experiencia y conocimientos	785	785.00	0.51	0.51
El tiempo para obtener ingresos o ganancias	289	1074.00	0.19	0.69
Los riesgos que conllevan	197	1271.00	0.13	0.82
La facilidad de producción	175	1446.00	0.11	0.94
Acceso a financiamiento	100	1546.00	0.06	1.00

Fuente: Elaboración propia.

## Personal de apoyo

Con respecto al personal que mantienen los emprendedores en su negocio bajo relación de dependencia, en su mayoría los entrevistados informaron de 1 persona (tabla 4.10). Seguramente debido a la situación económica, los mismos dueños son los que trabajan. Sin embargo, hay algunos emprendimientos que disponen de personal, como se puede apreciar en la figura 4.9), en la que de 5 a 10 personas requeridas representan aproximadamente un 40 % de las respuestas. Esto deja en claro que existen emprendimientos muy consolidados. En la muestra, entre 1 y 10 personas con relación de dependencia, representan aproximadamente el 99 % de las respuestas.

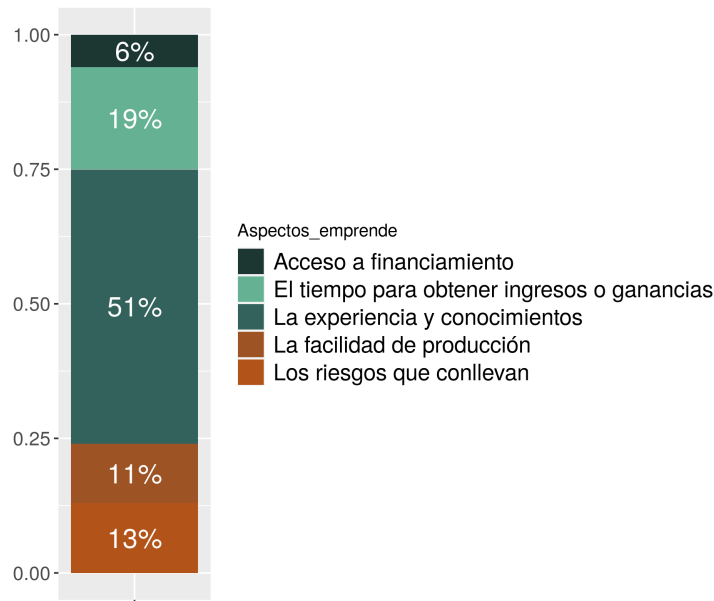


Figura 4.6: Aspectos considerados para el emprendimiento.  
Fuente: Elaboración propia.

Tabla 4.8: Distribución de frecuencias para la variable Problemas

Problemas	f	F	rf	rF
Competencia	937	937.00	0.61	0.61
Costo de materia prima	247	1184.00	0.16	0.77
Costo de mano de obra	129	1313.00	0.08	0.85
Costo de transporte	125	1438.00	0.08	0.93
Ninguna	80	1518.00	0.05	0.98
Otra	27	1545.00	0.02	1.00
Problemas familiares	1	1546.00	0.00	1.00

Fuente: Elaboración propia.

### Cantones objeto de la investigación

Se puede observar en la figura 4.10 los cinco cantones que son el objetivo de la investigación. Ellos se seleccionaron por la similitud de las actividades que mantienen. Esto es muy importante por lo cual, previamente se realizó un estudio en el que se identificó cuáles son sus actividades, mediante el código de Clasificación Industrial Internacional Uniforme (CIIU) que utiliza el Servicio de Rentas Internas (SRI) del Ecuador. Se observa que el cantón Patate es el que tiene más emprendedores que lograron ser entrevistados (ver tabla 4.11. En el margen de la tabla también se observa el total de 1546 emprendimientos con los que se pudo trabajar.

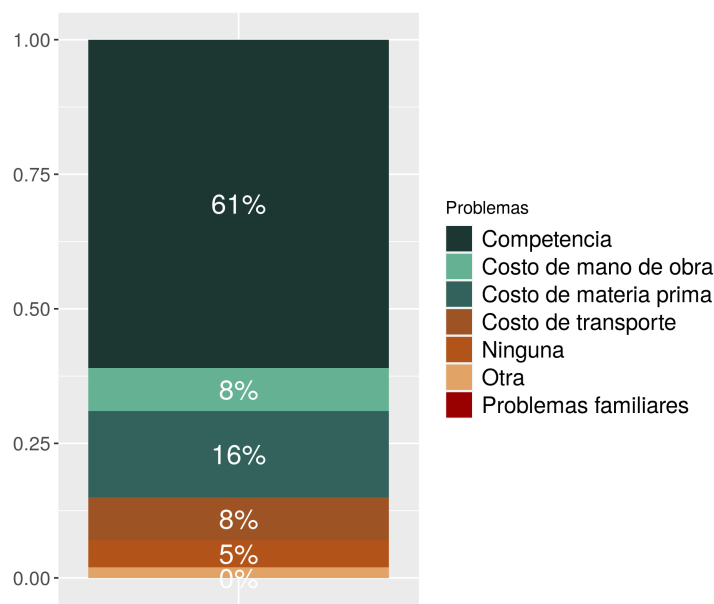


Figura 4.7: Problemas para el desarrollo del emprendimiento.  
Fuente: Elaboración propia.

Tabla 4.9: Distribución de frecuencias para la variable Publicidad

Publicidad	f	F	rf	rF
Redes sociales	533	533.00	0.34	0.34
Radio	512	1045.00	0.33	0.68
Otros	228	1273.00	0.15	0.82
Prensa escrita	149	1422.00	0.10	0.92
Páginas web	109	1531.00	0.07	0.99
Ninguno	15	1546.00	0.01	1.00

Fuente: Elaboración propia.

## Emprendimientos suspendidos

Claramente se puede observar que el número de emprendimientos suspendidos es pequeño con relación a aquellos que mantienen su actividad. Los emprendedores que manifiestan haber suspendido sus emprendimientos, se refieren a negocios que por diferentes problemas, notifican al SRI para que se registre que están suspendidos de manera temporal. Estos no han cerrado definitivamente, pudiendo así en un futuro retomar su actividad. Esto se puede observar en la figura 4.11 y la tabla 4.12, en las que se observa que existe un 75.23 % que mantienen su actividad versus el 24.77 % de los que se informa que están suspendidos temporalmente.

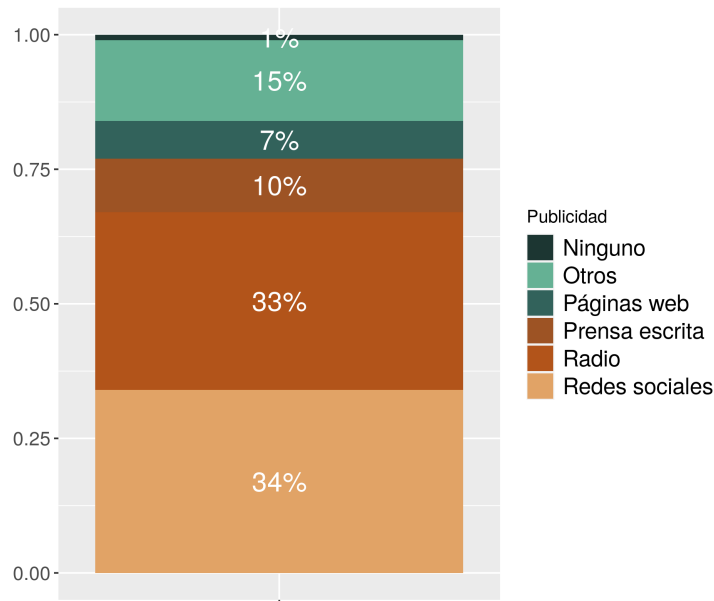


Figura 4.8: Tipo de publicidad para el emprendimiento.  
Fuente: Elaboración propia.

Tabla 4.10: Distribución de frecuencias para la variable Personal

Personal	f	F	rf	rF
1 persona	936	936.00	0.61	0.61
5 a 10 personas	582	1518.00	0.38	0.98
11 a 20 personas	23	1541.00	0.01	1.00
21 personas o mas	5	1546.00	0.00	1.00

Fuente: Elaboración propia.

## Emprendimientos cerrados

La última variable que se describe es la seleccionada como variable respuesta del estudio. En la figura 4.12, se observa que el número de emprendimientos cerrados definitivamente es mínimo (apenas un 10.61%). Los emprendimientos que mantienen su actividad representan un 89.39% de la muestra. Llama la atención que la proporción de emprendimientos cerrados es considerablemente inferior a la de suspendidos, mostrando que no todo emprendimiento suspendido necesariamente termina cerrando.

Los emprendimientos cerrados son negocios que, por diferentes problemas, como por ejemplo demasiado personal bajo relación de dependencia, falta de presupuesto, entre otros, tuvieron que cerrar. El cierre se notifica al SRI, haciendo caer directamente el registro de emprendimientos productivos en el país.

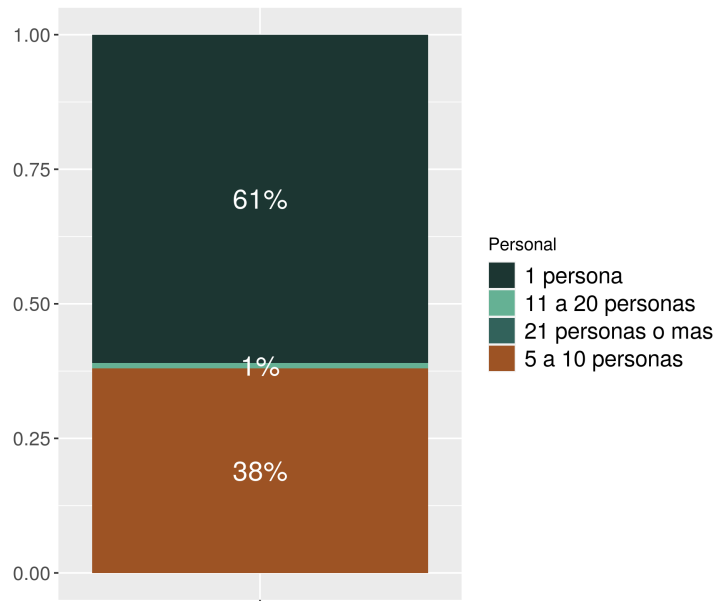


Figura 4.9: Personal de apoyo en relación de dependencia.  
Fuente: Elaboración propia.

Tabla 4.11: Distribución de frecuencias para la variable Cantón

Cantón	f	F	rf	rF
PATATE	426	426.00	0.28	0.28
QUERO	359	785.00	0.23	0.51
CEVALLOS	306	1091.00	0.20	0.71
TISALEO	303	1394.00	0.20	0.90
MOCHA	152	1546.00	0.10	1.00

Fuente: Elaboración propia.

#### 4.1.2. Análisis bidimensional

##### Distribución conjunta por edad y género

La tabla 4.14 contiene la distribución conjunta por edad y género. Claramente las clases con mayor representación en la muestra son el género masculino con edades comprendidas entre 39 y 59 años. La figura 4.13 muestra los diagramas de caja para ambas variables, considerando ahora la edad como variable entera.

En la figura 4.13 se notan claramente algunos *outliers* con relación a los emprendedores, tanto en el género masculino como femenino. Esto puede interpretarse como que cualquier persona, de cualquier edad, puede convertirse en un emprendedor. Hay datos (poco frecuentes) de emprendedores con 89 años y de la misma forma, de 19 años, no obstante la variabilidad de la edad es similar dentro de los géneros. Además, se puede observar en la figura, que la edad es ligeramente supe-

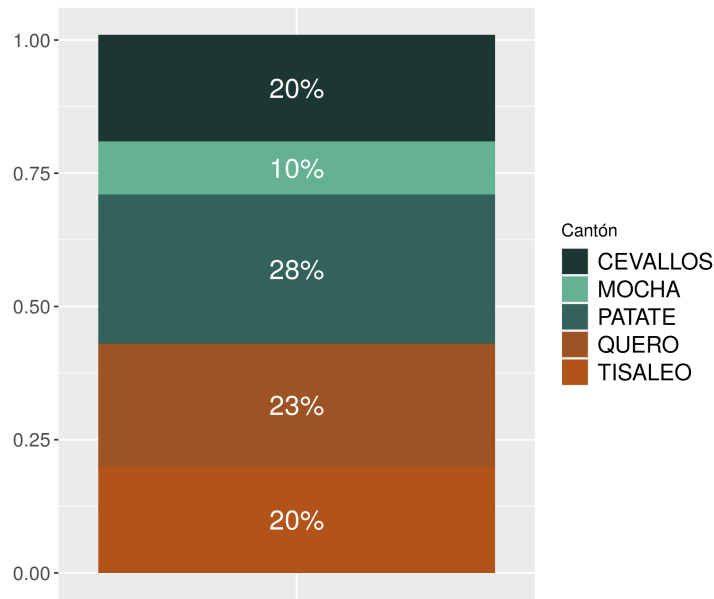


Figura 4.10: Cantones objeto de la investigación.  
Fuente: Elaboración propia.

Tabla 4.12: Distribución de frecuencias para la variable Suspendido

Suspendido	f	F	rf	rF
No	1163	1163.00	0.75	0.75
Si	383	1546.00	0.25	1.00

Fuente: Elaboración propia.

rior en los hombres que en las mujeres, aunque no se esperaría que la diferencia fuese significativa, ya que los cuerpos de ambos diagramas se solapan en sentido horizontal.

### Distribución conjunta por edad e instrucción

La tabla 4.15 contiene la distribución conjunta por edad y nivel de instrucción. Las clases con mayor representación en la muestra son la instrucción primaria y secundaria con edades comprendidas entre 39 y 59 años. La figura 4.14 muestra los diagramas de caja para ambas variables.

En la figura 4.14 se notan algunos *outliers* con relación a los emprendedores, para el nivel de instrucción primaria y secundaria, considerablemente más frecuentes en la muestra que los demás. La variabilidad de la edad es similar dentro de los niveles primaria, secundaria y tercer nivel, y más alta para la categoría sin instrucción. Se puede observar en la figura, que la edad es considerablemente superior entre los entrevistados sin instrucción e inferior entre los entrevistados con instrucción



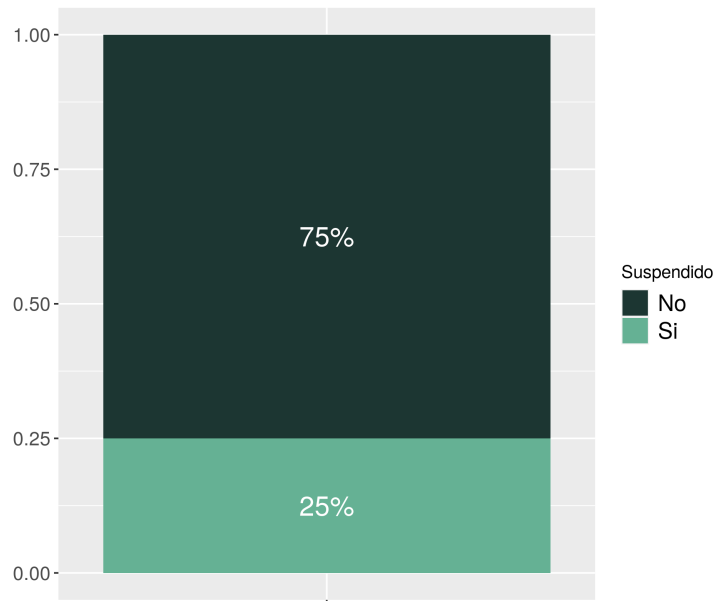


Figura 4.11: Emprendimientos suspendidos.

Fuente: Elaboración propia.

Tabla 4.13: Distribución de frecuencias para la variable Cerrado

Cerrado	f	F	rf	rF
No	1382	1382.00	0.89	0.89
Si	164	1546.00	0.11	1.00

Fuente: Elaboración propia.

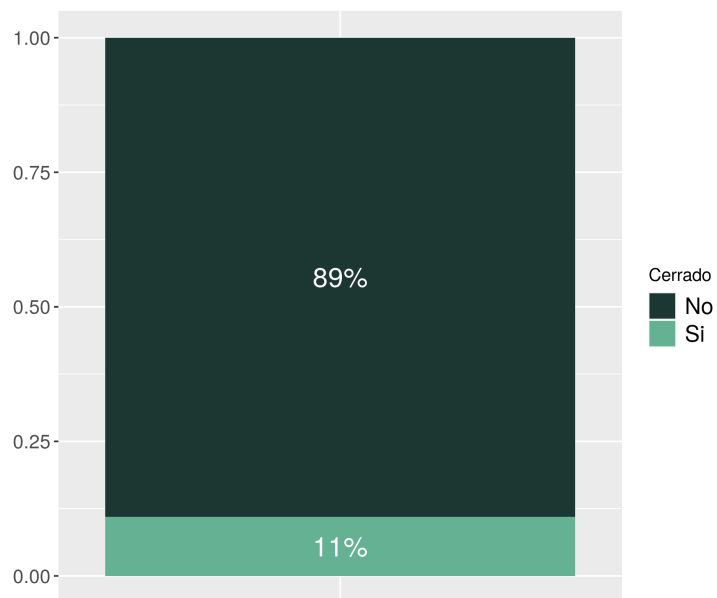


Figura 4.12: Emprendimientos cerrados.

Fuente: Elaboración propia.

Tabla 4.14: Distribución conjunta por edad y género

Edad	Género					
	Femenino		Masculino		Total	
	Nro.	%	Nro.	%	Nro.	%
[19,29)	24	1.55	25	1.62	49	3.17
[29,39)	145	9.38	146	9.44	291	18.82
[39,49)	208	13.45	336	21.73	544	35.19
[49,59)	152	9.83	329	21.28	481	31.11
[59,69)	42	2.72	97	6.27	139	8.99
[69,79)	15	0.97	16	1.03	31	2.01
[79,Inf)	1	0.06	10	0.65	11	0.71
Total	587	37.97	959	62.03	1546	100.00

Fuente: Elaboración propia.

Tabla 4.15: Distribución conjunta por Edad e Instrucción

Edad	Nivel de instrucción									
	Primaria		Secundaria		Sin Instrucción		Tercer Nivel		Total	
	Nro.	%	Nro.	%	Nro.	%	Nro.	%	Nro.	%
[19,29)	14	0.91	24	1.55	0	0.00	11	0.71	49	3.17
[29,39)	80	5.17	180	11.64	0	0.00	31	2.01	291	18.82
[39,49)	248	16.04	260	16.82	3	0.19	33	2.13	544	35.19
[49,59)	271	17.53	185	11.97	6	0.39	19	1.23	481	31.11
[59,69)	110	7.12	24	1.55	3	0.19	2	0.13	139	8.99
[69,79)	25	1.62	2	0.13	3	0.19	1	0.06	31	2.01
[79,Inf)	7	0.45	3	0.19	1	0.06	0	0.00	11	0.71
Total	755	48.84	678	43.86	16	1.03	97	6.27	1546	100.00

Fuente: Elaboración propia.

de tercer nivel. Esto indica que actualmente los emprendedores toman en cuenta la capacitación y el nivel de estudio, en mayor medida que tiempo atrás, cuando también pueden no haber existido las oportunidades educativas del presente.

### Distribución conjunta por edad y cantón

La tabla 4.16 contiene la distribución conjunta por edad y cantón de procedencia de los entrevistados. Existen en todos los cantones ciertos valores atípicos que se relacionan con emprendedores en una edad entre 70 y 89 años aproximadamente. Sólo en el cantón Cevallos se observa un dato por debajo de los 20 años, esto significa que hay un emprendedor muy joven. La figura 4.15 muestra los diagramas de caja para ambas variables.

En la figura 4.15. La variabilidad de la edad es muy similar dentro de los cantones, demostrando bastante uniformidad en la selección de la muestra original.

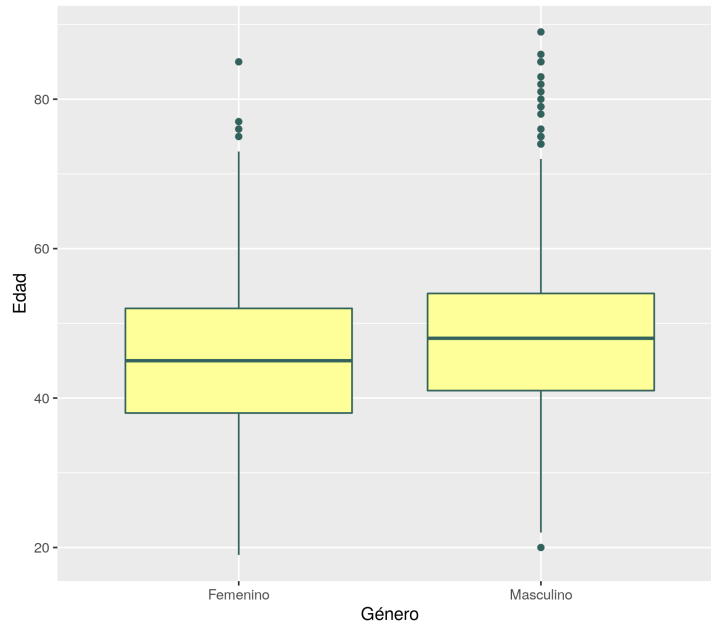


Figura 4.13: Diagrama de caja de la edad y género.  
Fuente: Elaboración propia.

Tabla 4.16: Distribución conjunta por Edad y Cantón

Edad	Cantón										Total	
	Cevallos		Mocha		Patate		Quero		Tisaleo			
	Nro.	%	Nro.	%	Nro.	%	Nro.	%	Nro.	%		
[19,29)	6	0.39	4	0.26	13	0.84	17	1.10	9	0.58	49	3.17
[29,39)	46	2.98	34	2.20	94	6.08	60	3.88	57	3.69	291	18.82
[39,49)	128	8.28	52	3.36	153	9.90	107	6.92	104	6.73	544	35.19
[49,59)	106	6.86	38	2.46	120	7.76	128	8.28	89	5.76	481	31.11
[59,69)	13	0.84	16	1.03	41	2.65	35	2.26	34	2.20	139	8.99
[69,79)	4	0.26	5	0.32	3	0.19	11	0.71	8	0.52	31	2.01
[79,Inf)	3	0.19	3	0.19	2	0.13	1	0.06	2	0.13	11	0.71
Total	306	19.79	152	9.83	426	27.55	359	23.22	303	19.60	1546	100.00

Fuente: Elaboración propia.

## 4.2. El modelo de regresión logística

Esta sección es el centro del trabajo. Explora los datos buscando el mejor modelo de regresión logística posible, considerando la variable “cerrado” como la respuesta. Para este caso, el valor ‘cerrado = Si’ se considera el éxito teórico (valor  $Y = 1$ ) y a el se refiere la probabilidad  $p$  de éxito modelada mediante el logit. El proceso de selección acude al método *step* ya que se prueba con todas las variables y se conservan sólo aquellas que resulten significativas. Este método procura cumplir con el principio de parsimonia, que consiste en seleccionar aquel modelo que tenga el menor número de parámetros posibles.

Debido a la cantidad de variables explicativas y el número de niveles que tiene cada una, los modelos empleados en este trabajo no consideran las interacciones,

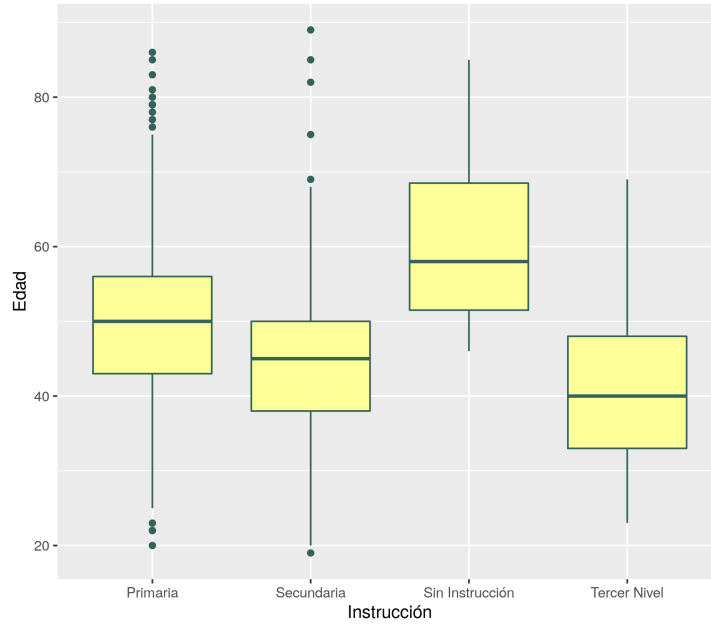


Figura 4.14: Diagrama de caja de la edad y el nivel de instrucción.  
Fuente: Elaboración propia.

así que se trata de modelos puramente aditivos.

El primer modelo que se prueba es el modelo completo, es decir, aquel que incluye todas las variables explicativas con que se cuenta. La formulación en términos generales es la siguiente:

$$\begin{aligned}
 \text{logit}(p) = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\
 & + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 \\
 & + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12}
 \end{aligned} \tag{4.1}$$

En (4.1), las variables explicativas tienen la correspondencia que se muestra en la tabla 4.17.  $\beta_0$  y  $\beta_4$  son escalares, el primero por tratarse del intercepto de la forma lineal y el segundo, por corresponder a una variable de tipo cuantitativo (Edad). Las restantes variables explicativas y sus coeficientes, son en realidad vectores, pues corresponde un coeficiente a cada nivel del factor que le acompaña.

Por ejemplo, Instrucción es un factor a 4 niveles (Primaria [31], Secundaria [32], Sin Instrucción [33], Tercer Nivel [34]), de forma que sus coeficientes asociados son  $\beta_3 = [\beta_{31} \ \beta_{32} \ \beta_{33} \ \beta_{34}]$ , correspondientes al vector de variables  $\mathbf{x}_3 = [x_{31} \ x_{32} \ x_{33} \ x_{34}]^T$ .

Para las variables cualitativas (todas, excepto la edad), su primer nivel se considera el nivel de referencia. Esto significa que se le impone una restricción de

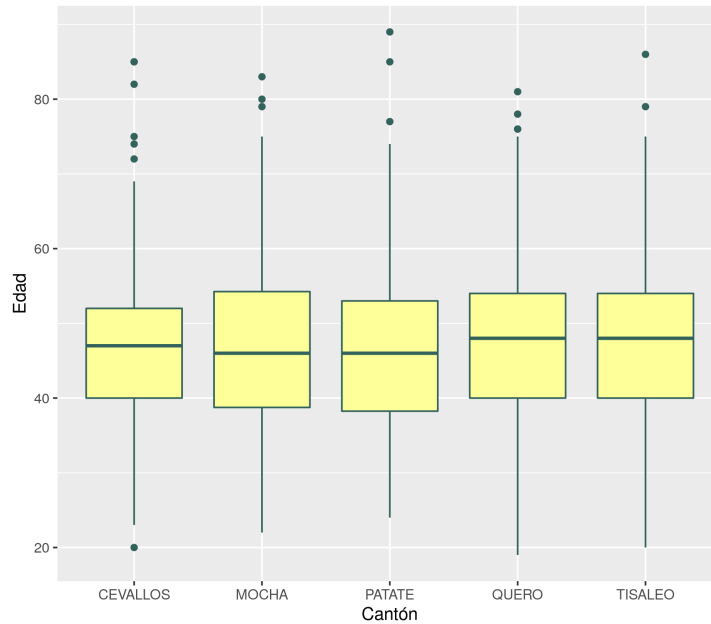


Figura 4.15: Diagrama de caja de la edad y el cantón.  
Fuente: Elaboración propia.

nulidad  $\beta_{i1} = 0, \forall i \neq 4$ . En consecuencia, la presencia del nivel de referencia en el modelo, se produce en la ausencia de los demás parámetros de la variable.

### 4.2.1. Las tablas ANOVA

Para el intercepto y cada variable, las tablas 4.18 y 4.19 muestran el análisis de la varianza con las estimaciones y las pruebas de hipótesis de nulidad de los parámetros. Normalmente todas las filas de este tipo de tablas se muestran juntas, sin embargo aquí se ha decidido separarlas en dos grupos en vista de su extensión.

Se han enfatizado los parámetros que resultan significativos considerando un nivel de confianza del 90%. Las variables que estadísticamente son significativas son aquellas en las que al menos uno de sus niveles resulta significativo. Estas son: Instrucción, Tiempo\_emprende, Suspendido, Personal y Cantón.

Para el resto de las variables explicativas, la muestra no contiene evidencia suficiente para rechazar la hipótesis de nulidad de sus parámetros en el modelo, por consiguiente, no parecen afectar el resultado de la variable respuesta (“cierre”). Esto no quiere decir que no sean variables importantes, es sólo que no muestran diferencias estadísticamente apreciables que expliquen el cierre definitivo de un emprendimiento productivo.

El procedimiento de ajuste `glm()` de **R** detecta un problema de subdispersión y

Tabla 4.17: Correspondencia entre las variables del archivo de datos y las del modelo

Variable de datos	Variable en el modelo	Parámetro asociado
Presupuesto	$\mathbf{x}_1$	$\beta_1$
Actividad	$\mathbf{x}_2$	$\beta_2$
Instrucción	$\mathbf{x}_3$	$\beta_3$
Edad	$x_4$	$\beta_4$
Género	$\mathbf{x}_5$	$\beta_5$
Tiempo_emprende	$\mathbf{x}_6$	$\beta_6$
Aspectos_emprende	$\mathbf{x}_7$	$\beta_7$
Problemas	$\mathbf{x}_8$	$\beta_8$
Publicidad	$\mathbf{x}_9$	$\beta_9$
Suspendido	$\mathbf{x}_{10}$	$\beta_{10}$
Personal	$\mathbf{x}_{11}$	$\beta_{11}$
Cantón	$\mathbf{x}_{12}$	$\beta_{12}$

Fuente: Elaboración propia.

lo corrige fijando el parámetro  $\phi = 1,450152$  (en lugar de 1, como cabría esperar en la situación ideal). El *deviance* del modelo nulo es 1045.84 sobre 1545 grados de libertad, mientras que el *deviance* residual (aquel que representa al error) es de 540.16 sobre 1490 grados de libertad.

La prueba de bondad del ajuste del *deviance* produce un *p*-valor muy cerca de 1, con lo cual debe considerarse que el modelo ajusta apropiadamente los datos.

#### 4.2.2. El modelo seleccionado

Analizadas todas las posibles variables, el modelo seleccionado se formula considerando sólo aquellas que resultaron significativas, esto es:

$$\text{logit}(p) = \beta_0 + \beta_3 \mathbf{x}_3 + \beta_6 \mathbf{x}_6 + \beta_{10} \mathbf{x}_{10} + \beta_{11} \mathbf{x}_{11} + \beta_{12} \mathbf{x}_{12} \quad (4.2)$$

La tabla 4.20 muestra los coeficientes y las pruebas de hipótesis para el modelo seleccionado. Nuevamente se resaltan los coeficientes que resultan significativos al 90 % de confianza.

Nótese de la tabla 4.20 que aunque no todos los coeficientes son significativos, al menos uno de cada factor lo es, indicando que todos los factores en consideración contribuyen a explicar el cierre del emprendimiento.

Tabla 4.18: Análisis de varianza

Coefficiente	Estimación	Error Est.	valor $t$	Pr(>   $t$  )
(Intercepto)	-1.54	1.87	-0.82	0.41
Presupuesto301 - 600	-0.26	0.44	-0.59	0.56
Presupuesto601 - 900	-0.21	0.46	-0.46	0.64
PresupuestoMás de 900	-0.36	0.46	-0.80	0.43
ActividadCría de aves de corral	0.11	1.57	0.07	0.95
ActividadCría de ganado y bovino	-0.49	1.15	-0.43	0.67
ActividadCultivo de plantas no perennes	0.09	1.06	0.09	0.93
ActividadCultivo de plantas perennes	0.21	1.08	0.19	0.85
ActividadElaboración de alimentos preparados	1.39	1.58	0.88	0.38
ActividadElaboración de bebidas	-16.06	1530.33	-0.01	0.99
ActividadElaboración de especias salsas	-11.80	4764.12	-0.00	1.00
ActividadElaboración de otros productos	-1.30	1.71	-0.76	0.45
ActividadElaboración de productos lácteos	-1.72	1.72	-1.00	0.32
ActividadElaboración y Conservación	-12.20	2112.73	-0.01	1.00
ActividadFabricación de calzado	-0.48	1.15	-0.42	0.67
ActividadFabricación de diversos productos de madera	-0.15	1.31	-0.12	0.91
ActividadFabricación de prendas de vestir	-0.42	1.23	-0.34	0.74
ActividadFabricación y comercialización de artículos de construcción	-11.55	4764.12	-0.00	1.00
ActividadFabricación y comercialización de otros productos	0.16	1.24	0.13	0.90
ActividadGanadería	-0.42	1.22	-0.34	0.73
ActividadOtro actividad en general	0.21	1.15	0.18	0.86
ActividadProducción de carne y de productos carnicos	-15.57	1860.81	-0.01	0.99
ActividadTejedura de Productos Textiles	-15.89	1177.91	-0.01	0.99
ActividadTropicales y Subtropicales	-11.97	4764.12	-0.00	1.00
ActividadVenta al por mayor de materiales para la construcción, artículos de ferretería, equipo, materiales de fontanería o gasfitería	-0.58	1.48	-0.39	0.69

Fuente: Elaboración propia.

Ahora, el procedimiento de ajuste `glm()` de **R** detecta un problema de subdispersión y lo corrige fijando el parámetro  $\phi = 1,452881$ . El *deviance* del modelo nulo es 1045.84 sobre 1545 grados de libertad, igual que antes, como cabe esperar, mientras que el *deviance* residual es de 584.54 sobre 1531 grados de libertad. Es de resaltar que el *deviance* residual bajó muy poco del modelo con todas las variables, al modelo seleccionado, reforzando la decisión de excluir las variables no significativas, pues prácticamente no contribuyen al estudio de la variable respuesta.

La prueba de bondad del ajuste del *deviance* produce un  $p$ -valor también muy cerca de 1, con lo cual debe considerarse que el modelo seleccionado ajusta apropiadamente los datos.

Este resultado da respuesta clara a las dos primeras interrogantes que motivan la investigación:

- A la pregunta ¿es o no posible explicar el éxito y el fracaso de un emprendi-

Tabla 4.19: Análisis de varianza (Cont.)

Coefficiente	Estimación	Error Est.	valor $t$	Pr(>   $t$  )
InstrucciónSecundaria	-0.58	0.32	-1.82	<b>0.07</b>
InstrucciónSin Instrucción	-0.16	1.46	-0.11	0.91
InstrucciónTercer Nivel	-0.71	0.78	-0.91	0.37
Edad	0.01	0.01	1.04	0.30
GéneroMasculino	-0.39	0.30	-1.31	0.19
Tiempo_emprendeHasta 2 años	-0.36	0.45	-0.80	0.42
Tiempo_emprendeHasta 3 años	-1.36	0.50	-2.74	<b>0.01</b>
Tiempo_emprendeHasta 4 años en adelante	-2.26	0.46	-4.92	<b>0.00</b>
Aspectos_emprendeEl tiempo para obtener ingresos o ganancias	-0.08	0.68	-0.12	0.90
Aspectos_emprendeLa experiencia y conocimientos	-0.24	0.63	-0.38	0.70
Aspectos_emprendeLa facilidad de producción	0.35	0.68	0.52	0.60
Aspectos_emprendeLos riesgos que conllevan	-0.19	0.73	-0.26	0.80
ProblemasCosto de mano de obra	-0.85	0.58	-1.48	0.14
ProblemasCosto de materia prima	0.05	0.41	0.11	0.91
ProblemasCosto de transporte	-0.44	0.53	-0.84	0.40
ProblemasNinguna	0.67	0.72	0.93	0.35
ProblemasOtra	-0.09	1.50	-0.06	0.95
ProblemasProblemas familiares	-14.84	4764.12	-0.00	1.00
PublicidadOtros	-0.76	1.30	-0.59	0.56
PublicidadPáginas web	-0.68	1.36	-0.50	0.62
PublicidadPrensa escrita	-2.07	1.43	-1.45	0.15
PublicidadRadio	-0.67	1.26	-0.53	0.60
PublicidadRedes sociales	-0.56	1.29	-0.43	0.67
SuspendidoSi	3.24	0.34	9.64	<b>0.00</b>
Personal11 a 20 personas	1.06	1.28	0.83	0.41
Personal21 personas o mas	-11.73	1884.63	-0.01	1.00
Personal5 a 10 personas	0.92	0.31	3.02	<b>0.00</b>
CantónMOCHA	-0.42	0.72	-0.58	0.56
CantónPATATE	-0.21	0.45	-0.48	0.63
CantónQUERO	-0.68	0.48	-1.41	0.16
CantónTISALEO	-1.50	0.54	-2.80	<b>0.01</b>

Fuente: Elaboración propia.

miento productivo de la provincia de Tungurahua, utilizando las variables: actividad, instrucción, edad, género, tiempo del emprendimiento, aspectos, problemas, publicidad, personal, cantón, si ha sido suspendido, o cerrado?, la respuesta es afirmativa, aunque sólo un subconjunto de las variables de interés, aportan a la explicación. Estas son: instrucción, tiempo del emprendimiento, personal, cantón y si ha sido suspendido.

- A la pregunta ¿es factible ajustar un modelo matemático/estadístico que permita estudiar este fenómeno?, la respuesta también es afirmativa. La ecuación (4.2) representa un modelo estadístico que ajusta apropiadamente los datos y permite estudiar relaciones entre variables que contribuyen a explicar el fenómeno del fracaso de los emprendimientos productivos.

En cuanto a las hipótesis propuestas, con este resultado se pueden confirmar o negar, como sigue:



Tabla 4.20: Análisis de varianza para el modelo seleccionado

	Coficiente	Estimación	Error Est.	valor $t$	Pr( $>  t $ )
	(Intercepto)	-2.20	0.51	-4.32	<b>0.00</b>
	InstrucciónSecundaria	-0.74	0.28	-2.66	<b>0.01</b>
	InstrucciónSin Instrucción	-0.07	1.39	-0.05	0.96
	InstrucciónTercer Nivel	-0.85	0.71	-1.20	0.23
	Tiempo_emprendeHasta 2 años	-0.26	0.40	-0.65	0.51
	Tiempo_emprendeHasta 3 años	-1.43	0.45	-3.19	<b>0.00</b>
Tiempo_emprendeHasta 4 años en adelante		-2.18	0.41	-5.35	<b>0.00</b>
	SuspendidoSi	3.16	0.31	10.07	<b>0.00</b>
	Personal11 a 20 personas	0.72	1.23	0.58	0.56
	Personal21 personas o mas	-11.29	756.16	-0.01	0.99
	Personal5 a 10 personas	0.79	0.27	2.86	<b>0.00</b>
	CantónMOCHA	-0.74	0.64	-1.15	0.25
	CantónPATATE	-0.01	0.35	-0.04	0.97
	CantónQUERO	-0.65	0.41	-1.61	0.11
	CantónTISALEO	-1.71	0.48	-3.57	<b>0.00</b>

Fuente: Elaboración propia.

- La hipótesis: “Existen factores que explican el éxito de los emprendimientos productivos en los cantones de la Provincia de Tungurahua” debe ser **aceptada**. En efecto, se ha encontrado evidencia empírica suficiente para afirmar que existen factores que explican el éxito de un emprendimiento productivo (aunque lo estudiado es el fracaso, la conclusión se obtiene del complemento).
- La hipótesis: “La zona geográfica cantón en donde se desarrolla su actividad, actividad, instrucción, edad, género, tiempo emprendimiento, aspectos, problemas, publicidad, personal, suspendido, cerrado, contribuyen a explicar el éxito o el fracaso de los emprendimientos productivos” debe ser parcialmente **aceptada**. Sólo se ha encontrado evidencia empírica de la contribución a la explicación de las variables instrucción, tiempo del emprendimiento, personal, cantón y si ha sido suspendido. Para el resto de las variables en consideración la hipótesis debe ser **rechazada**.
- La hipótesis: “El presupuesto promedio incide de manera significativa en el éxito de los emprendimientos productivos” debe ser **rechazada**. No se ha encontrado evidencia empírica que sustente tal afirmación. Esto puede deberse a que los rangos considerados son muy bajos. Posiblemente en rangos de inversión más altos, la conclusión sería diferente.
- La hipótesis: “El nivel de estudio incide en el éxito de los emprendimientos productivos” debe ser **aceptada**. Efectivamente se ha encontrado evidencia empírica para sustentar dicha afirmación.
- La hipótesis: “El ciclo de vida del emprendedor (tiempo que lleva en su emprendimiento) influye significativamente en el éxito” también debe ser

**aceptada.** Se ha encontrado evidencia empírica para sustentar dicha afirmación.

- La hipótesis: “Los emprendimientos suspendidos temporalmente explican el fracaso de los emprendimientos” debe ser **aceptada**. Efectivamente se ha encontrado evidencia empírica para sustentar dicha afirmación.
- La hipótesis: “El personal con el que cuenta bajo relación de dependencia explica el fracaso de los emprendimientos productivos” debe ser **aceptada**. Se ha encontrado evidencia empírica para sustentar dicha afirmación.
- La hipótesis: “La publicidad, influye en el éxito de los emprendimientos” debe ser **rechazada**. No se ha encontrado evidencia empírica que la sustente. Esto puede deberse a los bajos rangos de inversión en publicidad que se desprenden del contexto de la muestra. Muy posiblemente la publicidad sí tenga un efecto notorio, en escalas de inversión mayores.

Queda por responder la tercera pregunta que motiva la investigación, es decir, ¿en qué sentido y magnitud las variables consideradas explicativas afectan el resultado de un emprendimiento productivo? Este es el objeto de lo que sigue.

La tabla 4.21 reproduce los coeficientes estimados por el modelo, su exponenciación y los intervalos del 90 % de confianza estimados para los coeficientes exponenciados (con excepción del nivel “21 personas o más” de la variable Personal, que resulta muy amplio para mostrarse). Debe recordarse que al exponenciar (4.2) se obtienen los valores de las posibilidades (entendidas como la probabilidad del éxito  $p$  -cierre en este caso- dividida entre la probabilidad del fracaso  $[1 - p]$  -no cierre en este caso-). Esto es:

$$\begin{aligned} \exp[\text{logit}(p)] &= \exp \left[ \log \left( \frac{p}{1-p} \right) \right] = \frac{p}{1-p} \\ &= \exp(\beta_0 + \beta_3 x_3 + \beta_6 x_6 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12}) \\ &= e^{\beta_0} e^{\beta_3 x_3} e^{\beta_6 x_6} e^{\beta_{10} x_{10}} e^{\beta_{11} x_{11}} e^{\beta_{12} x_{12}} \end{aligned}$$

Conociendo que  $e^{ab} = (e^a)^b$ , es claro que un incremento unitario en el valor de la variable explicativa  $x$ , tiene un efecto multiplicativo, por un factor  $e^\beta$ , sobre la respuesta en términos de posibilidades, considerando que todos los demás factores permanecen inalterados.

Tabla 4.21: Coeficientes e intervalos de confianza al 90 %

	Estimación	Exp	Inf90	Sup90
(Intercept)	-2.20	0.110592	0.047069	0.251916
InstrucciónSecundaria	-0.74	0.476686	0.299766	0.749570
InstrucciónSin Instrucción	-0.07	0.931403	0.043567	6.513931
InstrucciónTercer Nivel	-0.85	0.426275	0.119481	1.279519
Tiempo_emprendeHasta 2 años	-0.26	0.769619	0.396706	1.489506
Tiempo_emprendeHasta 3 años	-1.43	0.238279	0.112432	0.495260
Tiempo_emprendeHasta 4 años en adelante	-2.18	0.112774	0.057156	0.219097
SuspendidoSi	3.16	23.600416	14.375898	40.557881
Personal11 a 20 personas	0.72	2.047308	0.202927	12.517700
Personal21 personas o mas	-11.29	0.000012		
Personal5 a 10 personas	0.79	2.194856	1.399726	3.463930
CantónMOCHA	-0.74	0.477889	0.151862	1.293221
CantónPATATE	-0.01	0.987849	0.558836	1.748855
CantónQUERO	-0.65	0.519687	0.263456	1.009879
CantónTISALEO	-1.71	0.180447	0.079687	0.389071

Fuente: Elaboración propia.

### Las posibilidades de la instrucción

Primaria es el nivel de referencia para la variable Instrucción. Todo el análisis se refiere a este nivel de referencia. Entonces, si todos los demás parámetros permanecen inalterados:

- Contar con instrucción secundaria multiplica por 0.48 la posibilidad de que el emprendimiento fracase, con respecto a aquellos que sólo cuentan con instrucción primaria. Es decir, la posibilidad de fracaso disminuye.
- No contar con instrucción multiplica por 0.93 la posibilidad de que el emprendimiento fracase, con respecto a aquellos que cuentan con instrucción primaria. La posibilidad de fracaso disminuye levemente. De hecho, puede considerarse que se comporta de forma similar al nivel de instrucción primaria.
- Contar con instrucción de tercer nivel multiplica por 0.43 la posibilidad de que el emprendimiento fracase, con respecto a aquellos que sólo cuentan con instrucción primaria. La posibilidad de fracaso disminuye aún más que cuando se cuenta con educación secundaria.

Queda demostrado claramente que a mayor nivel de instrucción, menor posibilidad de fracaso del emprendimiento. Con instrucción de tercer nivel, la posibilidad del fracaso se reduce a menos de la mitad que cuando sólo se cuenta con instrucción primaria.

## Las posibilidades del tiempo del emprendimiento

Hasta 1 año es el nivel de referencia para la variable Tiempo\_emprende. Entonces, si todos los demás parámetros permanecen inalterados:

- Que el emprendimiento tenga hasta 2 años funcionando multiplica por 0.77 la posibilidad de que fracase, con respecto a aquellos que sólo tienen hasta 1 año funcionando. La posibilidad de fracaso disminuye.
- Que el emprendimiento tenga hasta 3 años funcionando multiplica por 0.24 la posibilidad de que fracase, con respecto a aquellos que sólo tienen hasta 1 año funcionando. La posibilidad de fracaso disminuye considerablemente, de hecho es una cuarta parte.
- Que el emprendimiento tenga hasta 4 años funcionando multiplica por 0.11 la posibilidad de que fracase, con respecto a aquellos que sólo tienen hasta 1 año funcionando. La posibilidad de fracaso disminuye dramáticamente. Es casi una décima parte de la posibilidad del nivel de referencia.

Queda demostrado entonces que a mayor tiempo de funcionamiento, menor posibilidad de fracaso del emprendimiento. Con más de tres años de operación, las posibilidades de fracaso se reducen a casi la décima parte que cuando sólo se tiene hasta 1 año de operación. Puede decirse que un emprendimiento que supere los 3 años de funcionamiento tiene una posibilidad muy baja de fracasar. Este dato debe alentar a los emprendedores a resistir e insistir con sus emprendimientos y puede orientar a los organismos crediticios como un elemento a considerar para otorgar líneas de crédito viables.

## Las posibilidades del personal

Una persona es el nivel de referencia para la variable Personal. Entonces, si todos los demás parámetros permanecen inalterados:

- Que el emprendimiento tenga hasta de 1 a 11 personas contratadas multiplica por 0.72 la posibilidad de que fracase, con respecto a aquellos que sólo tienen 1 persona contratada. La posibilidad de fracaso disminuye.
- Que el emprendimiento tenga hasta 21 personas contratadas o más multiplica por 0.000012 la posibilidad de que fracase, con respecto a aquellos que

sólo tienen 1 persona contratada. Esto es, prácticamente anula la posibilidad del fracaso del emprendimiento.

- Que el emprendimiento tenga de 5 a 10 personas contratadas multiplica por 0.79 la posibilidad de que fracase, con respecto a aquellos que sólo tienen 1 persona contratada. También disminuye la posibilidad de fracaso.

Queda demostrado entonces que a mayor número de personas que trabajan en el emprendimiento, menor posibilidad de fracaso tiene. Es notable que en la muestra, aquellos emprendimientos con 21 personas o más trabajando, tienen una posibilidad ínfima de fracasar. Este es otro dato para las organizaciones crediticias, pero también para los emprendedores: la fuerza de trabajo fortalece las posibilidades de éxito de un emprendimiento.

### **Las posibilidades del cantón**

Cevallos es el nivel de referencia para la variable Cantón. Entonces, si todos los demás parámetros permanecen inalterados:

- Que el emprendimiento se ubique en el cantón Mocha multiplica por 0.48 la posibilidad de que fracase, con respecto a aquellos que se ubican en el cantón Cevallos. La posibilidad de fracaso disminuye.
- Que el emprendimiento se ubique en el cantón Patate multiplica por 0.99 la posibilidad de que fracase, con respecto a aquellos que se ubican en el cantón Cevallos. La posibilidad de fracaso disminuye casi imperceptiblemente. Podría decirse que Cevallos y Patate son dos cantones de comportamiento muy similar frente a los emprendimientos.
- Que el emprendimiento se ubique en el cantón Quero multiplica por 0.52 la posibilidad de que fracase, con respecto a aquellos que se ubican en el cantón Cevallos. La posibilidad de fracaso disminuye.
- Que el emprendimiento se ubique en el cantón Tisaleo multiplica por 0.18 la posibilidad de que fracase, con respecto a aquellos que se ubican en el cantón Cevallos. La posibilidad de fracaso disminuye dramáticamente. Sin duda es el cantón que más favorece el emprendimiento. Valdría la pena estudiar en el futuro, con mayor profundidad, por qué este comportamiento tan diferenciado.

Se demuestra entonces que no todos los cantones otorgan las mismas posibilidades de éxito a un emprendimiento. Tisaleo parece otorgar condiciones muy superiores al resto de los cantones considerados en la muestra.

### Las posibilidades de si ha sido suspendido

Finalmente, el nivel de referencia de la variable Suspendido es No. Entonces, si todos los demás parámetros permanecen inalterados, que el emprendimiento haya sido suspendido multiplica por 23.6 la posibilidad de que fracase, con respecto a aquellos que no hayan sido suspendidos. En este caso, la posibilidad de fracasar aumenta. Es evidente que un emprendedor debe luchar al máximo para que su emprendimiento no se tenga que declarar como suspendido. Los emprendimientos suspendidos muy rara vez se recuperan y tienen posibilidades muy altas de terminar cerrando definitivamente.

## 4.3. El modelo seleccionado como predictor / clasificador

Por último se analiza la capacidad de clasificación / predicción del modelo seleccionado, empleando la curva ROC. Para poder utilizar un modelo de regresión logística como clasificador, en primer lugar deben producirse las predicciones del modelo sobre las probabilidades, como se definen en (2.8), sustituyendo los valores teóricos de  $\beta$  por sus estimadores  $\hat{\beta}$ . Para el modelo seleccionado estas vienen dadas por:

$$\hat{p} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_3 x_3 + \hat{\beta}_6 x_6 + \hat{\beta}_{10} x_{10} + \hat{\beta}_{11} x_{11} + \hat{\beta}_{12} x_{12})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_3 x_3 + \hat{\beta}_6 x_6 + \hat{\beta}_{10} x_{10} + \hat{\beta}_{11} x_{11} + \hat{\beta}_{12} x_{12})}$$

El procedimiento `glm()` de R predice las probabilidades de éxito para cada observación contenida en los datos (1546 registros), que en este caso es la probabilidad de “cierre”. Una vez obtenidas las predicciones de las probabilidades, debe definirse una función de discriminación. En este caso se emplea la función  $\hat{p} < p_0$  para clasificar Cerrado = No. Resta definir el mejor valor de  $p_0$ . La tabla 4.22 despliega las cuatro matrices de confusión, en términos porcentuales, para  $p_0 = 0,5; 0,4; 0,3; 0,2$ .

En la tabla 4.22 se aprecia que  $p_0 = 0,5$  es el valor que mejor clasifica los No-

Tabla 4.22: Matrices de confusión (%) para  $p_0 = 0,5; 0,4; 0,3; 0,2$

Cierre	Pronóstico							
	$p_0 = 0,5$		$p_0 = 0,4$		$p_0 = 0,3$		$p_0 = 0,2$	
	No	Si	No	Si	No	Si	No	Si
No	97.03	2.97	96.16	3.84	93.49	6.51	90.09	9.91
Si	46.95	53.05	42.07	57.93	30.49	69.51	17.68	82.32

Fuente: Elaboración propia.

Cierres, pero es el que peor clasifica los Si-Cierres. Pasa lo contrario con  $p_0 = 0,2$ . Este valor discrimina mejor los Si-Cierres a costa de hacerlo peor con los No-Cierres.

Buscando un compromiso entre ambos extremos, pero en el entendido de que es más grave fallar en el Sí que en el No, se decide utilizar  $p_0 = 0,3$  en la función discriminante. Esta función clasifica correctamente el 69.51% de los cierres y el 93,49% de los no cierres.

La figura 5 muestra la curva ROC del clasificador seleccionado. Nótese que dicho clasificador se comporta muy bien, ya que la curva se aleja de la recta de referencia a 45 grados y se acerca bastante al ideal de una función en “L” invertida con el vértice a 90 grados. Un indicador numérico de la calidad del clasificador es el área bajo la curva ROC. En este caso, dicha área es de 0.92, bastante cerca del ideal unitario.

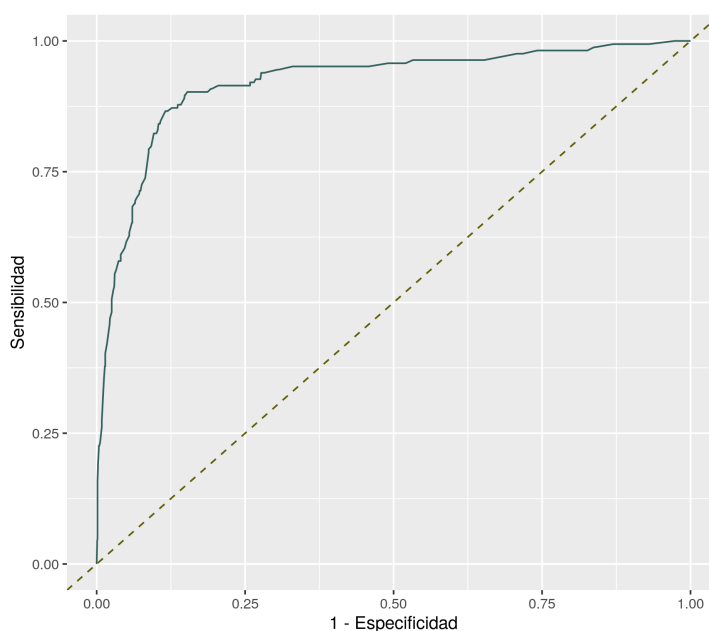


Figura 4.16: Curva ROC del modelo seleccionado ( $p_0 = 0,3$ ).

Fuente: Elaboración propia.

# Capítulo 5

## Conclusiones y recomendaciones

### 5.1. Conclusiones

Se comprueba la viabilidad del modelo propuesto de regresión logística, para la explicación y predicción del éxito o el fracaso de los emprendimientos productivos. La muestra consta de 1546 emprendimientos, distribuidos en cinco cantones, a saber: Cevallos, Quero, Mocha, Patate y Tisaleo. Se puede explicar el éxito o el fracaso de los emprendimientos productivos de la provincia de Tungurahua, utilizando las variables más significativas del modelo propuesto. Estas son: instrucción, tiempo del emprendimiento, personal, cantón y si ha sido suspendido. Las demás variables en consideración, no tienen incidencia significativa en la explicación del éxito o el fracaso de los emprendimientos productivos.

El modelo propuesto en (4.2) ajusta apropiadamente los datos y permite estudiar las relaciones causa/efecto entre las variables que explican el fenómeno del fracaso de los emprendimientos productivos.

Sobre las principales causas del éxito o el fracaso de los emprendimientos productivos en los cantones considerados, se puede concluir lo siguiente:

1. Se determina que, a mayor nivel de instrucción, menor posibilidad de fracaso del emprendimiento. Esto implica que existe evidencia suficiente en la muestra para sustentar que el nivel de estudio sí incide en el éxito de los emprendimientos productivos. El nivel de instrucción de tercer nivel reduce la posibilidad del fracaso a menos de la mitad, que el nivel de instrucción primaria.



2. Se demuestra que, a mayor tiempo que el emprendimiento está activo o en funcionamiento, menor es la posibilidad de fracaso. En un emprendimiento con más de tres años de operación, las posibilidades de fracaso se reducen a casi la décima parte que cuando sólo se tiene hasta 1 año de operación. Este resultado debe motivar la búsqueda de nuevas políticas de Estado por parte de entidades cooperantes, tales como: Consejo Nacional de Competencias, Honorable Gobierno Provincial De Tungurahua, CONAGOPARE, BANE-CUADOR, Cámara de Comercio de Ambato, para que otorguen líneas de crédito viables o bonos de desarrollo humano, tomando en consideración que el ciclo de vida del emprendimiento influye significativamente en el éxito.
3. El personal con el que se cuenta bajo relación de dependencia, es decir, el número de personas que trabajan en el emprendimiento productivo, explican su fracaso. A mayor número de personas que trabajan en el emprendimiento, menor es la posibilidad de fracaso. Este resultado deja claro que el talento humano es imprescindible como apoyo a los emprendimientos.
4. Los emprendimientos suspendidos temporalmente explican el fracaso. Un emprendimiento que haya sido suspendido, multiplica por 23.6 la posibilidad de que fracase, con relación a aquellos que no hayan sido suspendidos. Este resultado es importante porque los emprendimientos que están suspendidos, no necesariamente son un fracaso. Puede ser que suspendieron temporalmente por diversas razones. Sin embargo, los emprendimientos suspendidos tienen posibilidades muy altas de terminar cerrando definitivamente.
5. Con respecto a los cantones, no todos otorgan las mismas posibilidades de éxito a un emprendimiento. El cantón Tisaleo parece otorgar condiciones muy superiores al resto de los cantones considerados en la muestra.
6. Se determina que el presupuesto promedio y el tipo de publicidad, no inciden de manera significativa en el éxito de los emprendimientos productivos, es decir, no existe evidencia empírica que sustente tal afirmación, pudiendo deberse a que los rangos considerados son muy bajos.

Finalmente, en cuanto a líneas de investigación futuras, se tiene previsto aplicar técnicas de aprendizaje estadístico o de *Machine Learning*, una vez que se apliquen los instrumentos a los cantones que faltan por cubrir de la muestra original.

## 5.2. Recomendaciones

- Extender el presente estudio de regresión logística, considerando todos los cantones de la provincia de Tungurahua.
- Establecer convenios entre instituciones cooperantes para establecer políticas de Estado que ayuden a evitar el fracaso de los emprendimientos productivos, especialmente con líneas de crédito.
- Continuar con estudios futuros, empleando el presente trabajo como línea base, considerando los emprendimientos post-Covid 19. Estudiar los efectos de la pandemia en la situación de los emprendimientos y compararlos con el presente estudio.
- Analizar con mayor profundidad porque el presupuesto promedio anual no incide en el éxito de los emprendimientos, siendo que es una base el disponer de capital propio para el desarrollo del emprendimiento.
- Realizar estudios sobre las causas de los emprendimientos suspendidos temporalmente.
- Investigar con más detalle sobre las técnicas matemáticas y estadísticas alternativas de selección de variables como entradas al modelo.
- Estudiar modelos matemáticos alternativos que puedan ajustarse al conjunto de datos estudiado, obteniendo la base de datos del SRI actualizada al año 2021 mediante el código CIU.
- En futuras investigaciones, incluir más variables, tales como los ingresos, activos corrientes, deudas, entre otras. Es decir, incluir variables cuantitativas que permita realizar pronósticos en base a registros contables y que también den respuesta al éxito y fracaso de los emprendimientos de toda la provincia de Tungurahua.

# Bibliografía

- Alcaraz, R. (2011). *Emprendedor Éxito*. Ciudad de México: McGraw-Hill.
- Alvarez, C. and Noguera, M. and Urbano, D. (2012). Condicionantes del entorno y emprendimiento femenino. *Economía Industrial*(9).
- Bender, E. (2012). *An introduction to mathematical modeling*. Courier Corporation.
- Dobson, A., y Barnett, A. (2018). *An introduction to generalized linear models*. Taylor & Francis Group.
- Dym, C. (2004). *Principles of mathematical modeling*. Elsevier.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*(27, 8), 861 – 874. doi: <https://doi.org/10.1016/j.patrec.2005.10.010>
- Gómez, E., Lafuente, E., Vaillant, Y., y Gómez, L. (2015). El impacto diferenciado de la autoconfianza, los modelos de referencia y el miedo al fracaso sobre los jóvenes emprendedores. *Innovar*(18).
- Halabi, C., y Lussier, R. (2014). A model for predicting small firm performance increasing the probability of entrepreneurial success in Chile. *Journal of Small Business and Enterprise Development*(22).
- Heredia, J., Rodríguez, A., y Vilalta, J. (2014). Predicción del rendimiento en una asignatura empleando la regresión. *Estudios Pedagógicos*(18).
- Hernández, R., Fernández, C., y Baptista, P. (2014). *Metodología de la investigación*. México D.F.: McGraw/Hill.
- Hosmer, D. W., y Lemeshow, S. (2000). *Applied logistic regression* (2.<sup>a</sup> ed.). NY, EEUU: John Wiley & Sons.
- Hyder, S., y Lussier, R. (2016). Why Businesses Succeed or Fail:A Study on Small Businesses in Pakistan. *Journal of Entrepreneurship in Emerging Economies*(23).
- Kilkenny, M., Nalbarte, L., y Besser, T. (2013). Reciprocated community support and small town-small business success. *Entrepreneurship & Regional Development. An International Journal*(18).
- Laitinen, E. (2013). Financial and non-financial variables in predicting failure of small business reorganisation. *Accounting and Finance*(34).


- Lasio, V., Amaya, A., Zambrano, J., y Ordeñana, X. (2020). *Global entrepreneurship monitor*. Quito.
- Leary, M., y DeV Vaughn, M. (2009). Entrepreneurial team characteristics that influence the successful launch of a new venture. *Journal of Small Business and Enterprise Development*(6, 14).
- Lussier, R., y Pfeifer, S. (2001). A Crossnational Prediction Model for Business Succes. *Journal of Small Business Management*(12).
- Mavi, R., Mavi, N., y Goh, M. (2015). Modeling corporate entrepreneurship success with ANFIS. *Operational Research An International Journal*(26).
- Messina, M., y Hochsztain, E. (2015). Factores de éxito de un emprendimiento, Un estudio exploratorio con base en técnicas de data mining. *Dialnet*(10).
- Morales, J., Bustamante, A., Vargas, S., Pérez, N., y Sereno, O. (2015). Factores de éxito emprendedor en dos municipios de la montaña de Guerrero, México. *Nova Scientia*(20).
- Navarrete, E., y Sansore, E. (2011). El fracaso de las micro, pequeñas y medianas empresas en Quintana Roo, México: Un análisis multivariante. *Revista Internacional Administración y Finanzas*(13).
- Nelder, J. A., y Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Serie A*(135, 3), 370–384.
- Núñez, J. V., Peñalosa, V. L., y Armijos, E. J. (2015). Principales causas de mortalidad de las compañías de comercio en la región central del Ecuador. *E-Mercatoria*(21).
- Ponsot, E. (2011). *Estudio de la agrupación de niveles en el modelo logit* (phdthesis). Universidad de Los Andes, Instituto de Estadística Aplicada y Computación, Facultad de Ciencias Económicas y Sociales, Mérida, Venezuela.
- Ponsot, E., y Sinha, S. (2009). Un modelo de regresión logística del rendimiento en los estudios universitarios: Caso FACES-ULA. *Actualidad Contable FACES*(23).
- R Core Team. (2020). R: A language and environment for statistical computing [Manual de software informático]. Vienna, Austria. Descargado de <https://www.R-project.org/>
- Rencher, A., y Schaalje, B. (2008). *Linear models in statistics*. John Wiley & Sons.
- Romero, M., Haro, M., y Ramírez, E. (2017). El Estado Colombiano y el emprendimiento empresarial: éxito o fracaso de su programa clave. *Criterio Libre*(28).
- Silva, G., Freire, P., Lina, M., Tavares, N., y Rojas, A. (2016). Entrepreneurial potential and success in business: a study on elements of convergence and

explanation. *Revista de Administración Mackenzie*(28).


- Vainrub, R. (2009). *Una guía para emprendedores*. Ciudad de México: Pearson.
- Van Gelderen, M., Thurik, R., y Bosma, N. (2005). Success and Risk Factors in the Pre-Startup Phase. *Small Business Economics*(16).
- Yeh, C.-C., Chi, D.-J., y Hsu, M.-F. (2010). A hybrid approach of DEA, rough set and support vector machine business failure prediction. *Expert Systems with Applications*(37, 2), 1535–1541.

# Anexos

## Anexo 1: Instrumento de recolección de datos



UNIVERSIDAD TÉCNICA DE AMBATO  
FACULTAD DE CONTABILIDAD Y AUDITORÍA  
PROYECTO DE VINCULACIÓN



La siguiente encuesta tiene como objetivo examinar la sostenibilidad en los emprendimientos del sector productivo de Tungurahua.

**INDICACIONES**  
Seleccione la respuesta que usted crea pertinente según su criterio.

Género	Edad	Nivel de educación
Masculino		Primaria
Femenino		Secundaria
	Entre 20 -30 años	
	Entre 30 - 40 años	
	Entre 40 – más años	Tercer nivel
		Sin instrucción

Pertenece algún gremio o asociación

**1. ¿Cuál es la actividad principal de su emprendimiento?**

<input type="checkbox"/> Cultivo de verduras, hortalizas y tubérculos
<input type="checkbox"/> Cultivo de fruta
<input type="checkbox"/> Cría y reproducción de animales(ganado bovino, vacuno, cerdos, pollos y aves)
<input type="checkbox"/> Actividades agrícolas y de ganadería
<input type="checkbox"/> Producción, preparación, faenamiento y empackado de carne
<input type="checkbox"/> Fabricación de productos cármicos
<input type="checkbox"/> Producción de leche de vaca y huevos de ave de corral
<input type="checkbox"/> Elaboración de productos lácteos
<input type="checkbox"/> Elaboración de alimentos a base de cereal
<input type="checkbox"/> Elaboración y comercialización de productos alimenticios, dulces y chocolates
<input type="checkbox"/> Procesamiento de especias
<input type="checkbox"/> Fabricación y comercialización de alimentos preparados para animales
<input type="checkbox"/> Elaboración de bebidas alcohólicas y malteadas
<input type="checkbox"/> Fabricación de tejidos y sus derivados
<input type="checkbox"/> Fabricación, confección y comercialización de prendas de vestir
<input type="checkbox"/> Fabricación y comercialización de calzado
<input type="checkbox"/> Fabricación y comercialización de artículos de madera
<input type="checkbox"/> Fabricación y comercialización de artículos de construcción
<input type="checkbox"/> Elaboración y comercialización de productos (librerías, panaderías, cafeterías, bares, bisutería, juguetería, etc.)

**2. ¿Para llevar a cabo su emprendimiento que aspectos considero?**

<input type="checkbox"/> Los riesgos que conlleva
<input type="checkbox"/> El tiempo para obtener ingresos o ganancias
<input type="checkbox"/> La experiencia y conocimientos
<input type="checkbox"/> La facilidad de producción
<input type="checkbox"/> Acceso a financiamiento

Figura 1: Instrumento de recolección de datos. Página 1.

Fuente: Estudio de sostenibilidad de los emprendimientos productivos de la Provincia de Tungurahua, Ecuador, Frente 1 - Fase 2.

3. ¿Qué factores considera para realizar la producción o elaboración de su producto?

<input type="checkbox"/>	Calidad
<input type="checkbox"/>	Costo
<input type="checkbox"/>	Cantidad
<input type="checkbox"/>	Margen de utilidad
<input type="checkbox"/>	Necesidad si la opto como única fuente de ingreso o pérdida de empleo
<input type="checkbox"/>	Oportunidad

4. ¿Cuál es el presupuesto promedio anual que invierte en su actividad económica o emprendimiento?

<input type="checkbox"/>	\$0 - \$300
<input type="checkbox"/>	\$301 - \$600
<input type="checkbox"/>	\$601 - \$900
<input type="checkbox"/>	Más de \$901

5. ¿Qué tipo de publicidad considera que es necesario para la promoción de su producto?

<input type="checkbox"/>	Radio
<input type="checkbox"/>	Prensa escrita
<input type="checkbox"/>	Redes sociales
<input type="checkbox"/>	Página web
<input type="checkbox"/>	Otro .....

6. ¿Su actividad económica le genera utilidad?

<input type="checkbox"/>	Si
<input type="checkbox"/>	No

7. ¿A encontrado algún problema que le impide desarrollar de manera eficiente su actividad emprendedora?

<input type="checkbox"/>	Competencia
<input type="checkbox"/>	Costo de materia prima
<input type="checkbox"/>	Costos de transporte
<input type="checkbox"/>	Costo de mano de obra
<input type="checkbox"/>	Ninguna
<input type="checkbox"/>	Otra

8. ¿Su actividad económica está financiada por?

<input type="checkbox"/>	Ahorros personales
<input type="checkbox"/>	Préstamos a familiares o amigos
<input type="checkbox"/>	Programas del gobierno
<input type="checkbox"/>	Préstamo de instituciones financieras

Figura 2: Instrumento de recolección de datos. Página 2.  
Fuente: Estudio de sostenibilidad de los emprendimientos productivos de la Provincia de Tungurahua, Ecuador, Frente 1 - Fase 2.

9. ¿Ha recibido algún tipo de apoyo por parte de la autoridad gubernamental?

<input type="checkbox"/>	Si
<input type="checkbox"/>	No

10. ¿Considera usted que su actividad económica es sostenible dentro de los próximos 5 años?

<input type="checkbox"/>	Si
<input type="checkbox"/>	No

11. ¿Dentro de su actividad económica efectúa proyectos productivos?

<input type="checkbox"/>	Si
<input type="checkbox"/>	No

\* En caso de ser negativa su respuesta dirijase a la pregunta N° 12

\* En caso de ser positiva su respuesta dirijase a la pregunta N° 24

12. ¿Qué desea alcanzar con un Proyecto Productivo?

<input type="checkbox"/>	Sostenibilidad del emprendimiento en el tiempo
<input type="checkbox"/>	Relacionarse con los ciudadanos de una forma más afin
<input type="checkbox"/>	Generar rentabilidad económica y obtener ganancias
<input type="checkbox"/>	Producir bienes y servicios en un solo sector

13. ¿Qué factor considera usted para el desempeño de la productividad?

<input type="checkbox"/>	Motivación del personal
<input type="checkbox"/>	Talento humano
<input type="checkbox"/>	El capital
<input type="checkbox"/>	Canales de comercialización

14. ¿Qué factores considera que sea uno de los motivos para el cierre de la actividad económica?

<input type="checkbox"/>	Falta de financiamiento
<input type="checkbox"/>	La actividad económica no genera utilidad
<input type="checkbox"/>	No tener bases sólidas ni estrategias de negocios
<input type="checkbox"/>	Por productos sustitutos
<input type="checkbox"/>	Competencia
<input type="checkbox"/>	Estabilidad económica y política del país
<input type="checkbox"/>	Falta de apoyo de gestión del gobierno local

Figura 3: Instrumento de recolección de datos. Página 3.

Fuente: Estudio de sostenibilidad de los emprendimientos productivos de la Provincia de Tungurahua, Ecuador, Frente 1 - Fase 2.



**15. ¿Ha suspendido su negocio o actividad económica?**

Si
No

**16. ¿Ha cerrado definitivamente su negocio?**

Si
No

*\* En caso de ser negativa su respuesta dirijase a la pregunta N° 18.*

**17. ¿Cuándo realizo el trámite de cierre del negocio también lo hizo de su RUC?**

Si
No

**18. ¿Efectuó el trámite de cierre del RUC?**

Si
No

**19. ¿Le gustaría recibir un modelo de negocio que le permita tener un sistema de formación que le aporte a tener éxito?**

Si
No

**20. ¿Qué tiempo tiene ejecutando la actividad económica o emprendimiento?**

Hasta 1 año
Hasta 2 años
Hasta 3 años
De 4 años en adelante

**21. ¿La actividad económica que ejecuta cuenta con personal de apoyo en relación de dependencia?**

1 persona
5 a 10 personas
11 a 20 personas
Más de 21 personas

**22. ¿Identifique sus principales productos, cuál es su tiempo de producción, y lugar de comercialización?**

Producto	tiempo	Lugar de comercialización

Figura 4: Instrumento de recolección de datos. Página 4.  
Fuente: Estudio de sostenibilidad de los emprendimientos productivos de la Provincia de Tungurahua, Ecuador, Frente 1 - Fase 2.

23. ¿Señale la condición que le permitirá la Sostenibilidad del emprendimiento?

.....
-------

24. ¿Señale qué tipo de proyecto productivo realizó?

.....
-------

25. ¿Considera necesario recibir capacitación para mejorar su emprendimiento?

<input type="checkbox"/>	Si
<input type="checkbox"/>	No
<input type="checkbox"/>	Especifique el tema

Figura 5: Instrumento de recolección de datos. Página 5.  
Fuente: Estudio de sostenibilidad de los emprendimientos productivos de la Provincia de Tungurahua, Ecuador, Frente 1 - Fase 2.

## Anexo 2: Código R programado

### Funciones utilitarias

```
library(R6)
library(dplyr)
### Construcción de tablas de frecuencias
### Llamar con FreqTab$new()
FreqTab <- R6Class("FreqTab",
  public = list(
    data = NULL,
    dtype = NULL, # Tabla cruda. Otra cosa si quiero a partir
                  # de las frecuencias
    amplitude = NULL,
    name = NULL, # Nombre de la clase tabulada
    tab = NULL,
    iasf = TRUE, # TRUE: Integer as Factor, FALSE: Integer as Real
    fOrder = TRUE, # TRUE: frequency descending order
    initialize = function(data = NA, dtype=self$dtype,
                          iasf = self$iasf, amplitude = NA,
                          name=NA, fOrder = self$fOrder) {
      if (is.null(dtype)) {
        if (is.factor(data)||is.character(data)) {
          self$build_factor(data, name, fOrder)
        } else {
          if (is.integer(data)) {
            if (iasf == TRUE) {
              self$build_integer(data, name, fOrder)
            } else {
              self$build_real(data, amplitude, name)
            }
          } else {
            self$build_real(data, amplitude, name)
          }
        }
      } else {
        self$build_freq(data)
      }
    },
    build_freq = function(data=self$data) {
      self$tab <- data[,1:2]
      n_cat <- nrow(self$tab)
      total <- sum(self$tab$f)
      F <- c(rep(0, n_cat))
      F[1] <- self$tab$f[1]
      if (n_cat>1) {
```

```

        for (i in 2:n_cat) {
          F[i] <- F[i-1]+self$tab$f[i]
        }
      }
      self$tab <- cbind(self$tab , F=F,
                       rf=round(self$tab$f/total,2))
      self$tab <- cbind(self$tab, rF=round(self$tab$F/total,2))
      #self$n_classes <- n_cat
    },
    build_factor = function(data=self$data, name=self$name,
                           fOrder=self$fOrder) {
      if (fOrder==TRUE) {
        self$tab <- arrange(as.data.frame(table(data)),desc(Freq))
      } else {
        self$tab <- as.data.frame(table(data))
      }
      n_cat <- nrow(self$tab)
      total <- sum(self$tab$Freq)
      F <- c(rep(0, n_cat))
      F[1] <- self$tab$Freq[1]
      if (n_cat>1) {
        for (i in 2:n_cat) {
          F[i] <- F[i-1]+self$tab$Freq[i]
        }
      }
      self$tab <- cbind(self$tab , F, round(self$tab$Freq/total,2))
      self$tab <- cbind(self$tab, round(self$tab$F/total,2))
      colnames(self$tab) <- c(name, "f", "F", "rf", "rF")
      #self$n_classes <- n_cat
    },
    build_integer = function(data=self$data, name=self$name,
                             fOrder=self$fOrder) {
      if (fOrder==TRUE) {
        self$tab <- arrange(as.data.frame(table(data)),desc(Freq))
      } else {
        self$tab <- as.data.frame(table(data))
      }
      n_cat <- nrow(self$tab)
      total <- sum(self$tab$Freq)
      F <- c(rep(0, n_cat))
      F[1] <- self$tab$Freq[1]
      if (n_cat>1) {
        for (i in 2:n_cat) {
          F[i] <- F[i-1]+self$tab$Freq[i]
        }
      }
      self$tab <- cbind(self$tab, F, round(self$tab$Freq/total,2))

```

```

self$tab <- cbind(self$tab, round(self$tab$F/total,2))
colnames(self$tab) <- c(name, "f", "F", "rf", "rF")
},
build_real = function(data=self$data, amplitude=self$amplitude,
                      name=self$name) {
  range <- range(data)
  if (is.na(amplitude)) {
    n_classes <- trunc(1+3.322*log10(length(data)))
    a_classes <- (range[2]-range[1])/n_classes
  } else {
    n_classes <- ((range[2]-range[1])/amplitude) + 1
    a_classes <- amplitude
  }
  cuts <- seq(range[1], range[2], by=a_classes)
  cuts[n_classes] <- Inf
  data.cut <- cut(data, cuts, include.lowest=TRUE,
                 right=FALSE)
  self$tab <- as.data.frame(table(data.cut))
  total <- sum(self$tab$Freq)
  F <- c(rep(0, n_classes))
  F[1] <- self$tab$Freq[1]
  if (n_classes>1) {
    for (i in 2:n_classes) {
      F[i] <- F[i-1]+self$tab$Freq[i]
    }
  }
  self$tab <- cbind(self$tab, F, round(self$tab$Freq/total,2))
  self$tab <- cbind(self$tab, round(self$tab$F/total,2))
  colnames(self$tab) <- c(name, "f", "F", "rf", "rF")
  self$tab[,1] <- gsub('[', '|', self$tab[,1], fixed=TRUE)
  self$tab[,1] <- gsub(']', '|', self$tab[,1], fixed=TRUE)
},
print = function() {
  #res <- list(tab=self$tab, k=n_classes)
  return(self$tab)
}
)
)

# Tabla de contingencia sencilla para dos variables
# dat: data.frame, var1, var2: character()
tabCon2 <- function(dat, var1 = NULL, var2 = NULL) {
  formu <- paste0("~", var1, "+", var2)
  xt_f <- xtabs(formu, data = dat)
  xt_p <- prop.table(xt_f)
  xt_f <- addmargins(xt_f)
  xt_p <- addmargins(xt_p)
}

```

```

n <- nrow(xt_p)
m <- ncol(xt_p)
df <- data.frame(matrix(0, ncol = m*2, nrow = n))
nom_tmp <- c(levels(dat[, var2]), "Total")
nom_cols <- vector(length = 2*m)
j <- 1
for (i in 1:m) {
  nom_cols[j] <- paste0(nom_tmp[i], "_Nro")
  nom_cols[j+1] <- paste0(nom_tmp[i], "_Por")
  j <- j + 2
}
colnames(df) <- nom_cols
for (i in 1:n) {
  j <- 1
  for (k in 1:m) {
    df[i,j] <- xt_f[i,k]
    df[i,j+1] <- round(xt_p[i,k]*100, 2)
    j <- j + 2
  }
}
nom_tmp <- c(levels(dat[, var1]), "Total")
rownames(df) <- nom_tmp
deci <- c(0, rep(c(0,2),m))
return(xtable(df, digits=deci,
              caption = paste0("Distribución conjunta por ",
                                var1, " y ", var2),
              label = paste0("tab_", reempaza_esp(var1),
                              "_", reempaza_esp(var2))))
}

```

## Análisis unidimensional

```

library(ggplot2)
library(xtable)
source("MyLib.R")

# Lectura de datos
dat <- read.table("./Datos/emprende.csv",
                 header = TRUE, dec=".", sep = ";")
dat <- subset(dat, select = -c(X,X.1))

# Datos de las variables
info_vars <- data.frame(nombre=character(), clase = character(),
                       niveles = character())
n_vars <- ncol(dat)
nom_vars <- colnames(dat)

```

```

for (nom in nom_vars) {
  clase <- class(dat[, nom])
  if (clase == "factor") {
    niveles <- paste0(levels(dat[, nom]), collapse = ", ")
  } else {
    niveles <- "NA"
  }
  info_vars <- rbind(info_vars, cbind(nom, clase, niveles))
}
sink(paste0("./Tesis/Latex/Tablas/tb_var.tex"))
print(xtable(info_vars, caption="Variables y tipos",
             label = "tab_vars"),
      caption.placement="top",
      include.rownames=FALSE,
      size="\\fontsize{10pt}{11pt}\\selectfont")
sink(type="message")
sink()
#nom_vars <- "Actividad"

### Análisis univariante: Tabla de frecuencia y gráficos
for (nom in nom_vars) {
  clase <- class(dat[, nom])
  if (clase == "factor") {
    f <- FreqTab$new(dat[, nom],
                    name = nom)
    p <- ggplot(f$tab, aes(x="", y=rf, fill=get(nom))) +
      geom_bar(width = 1, stat = "identity",
              position = position_stack(reverse = FALSE)) +
      # coord_polar(theta="y") +
      geom_text(aes(label = paste0(round(rf*100, 2), "%")),
              position = position_stack(vjust = 0.5,
                                      reverse = FALSE),
              check_overlap = TRUE, size=6, color="white") +
      xlab(NULL) +
      ylab(NULL) +
      labs(fill = nom, size = 14) +
      theme(legend.text = element_text(size = 14),
            axis.text = element_text(size = 12)) +
      # legend.position="bottom"
      scale_fill_manual(values=pColores())
  } else {
    f <- FreqTab$new(dat[, nom],
                    name = nom,
                    iasf = FALSE,
                    fOrder = FALSE)
    p <- ggplot(dat, aes(x=get(nom))) +
      geom_histogram(color = pColores("named")["cMarron0"],

```

```

        fill = pColores("named")["cAmaC"],
        bins = 15) +
  theme(legend.text = element_text(size = 10)) +
  xlab(nom) +
  ylab("Frecuencias simples")
}
sink(paste0("./Tesis/Latex/Tablas/tb_", reempaza_esp(nom), ".tex"))
print(xtable(f$tab, label = paste0("tab_", reempaza_esp(nom)),
  caption = paste0("Distribución de frecuencias ",
    "para la variable ", nom)),
  caption.placement="top",
  include.rownames=FALSE,
  size="\\fontsize{10pt}{11pt}\\selectfont")
sink(type="message")
sink()
png(paste0("./Tesis/Latex/Figuras/fg_", reempaza_esp(nom), ".png"),
  width = 16, height = 14, units = "cm", res = 300)
print(p)
dev.off()
}

```

## Análisis bi-dimensional

```

library(ggplot2)
library(xtable)
library(stringr)

source("MyLib.R")

dat <- read.table("./Datos/emprende.csv",
  header = TRUE, dec=".", sep = ";")
dat <- subset(dat, select = -c(X,X.1))

## Edad y género
dat$Edad_recod <- recod_num(dat$Edad, amp = 10)

tabCon2(dat, "Edad_recod", "Género")

p <- ggplot(dat, aes(x=Género, y=Edad)) +
  geom_boxplot(color = pColores("named")["cAzul0"],
    fill = pColores("named")["cAmaC"])
png(paste0("./Tesis/Latex/Figuras/fg_Edad_Genero.png"),
  width = 16, height = 14, units = "cm", res = 300)
p
dev.off()

```



```

# Edad e instrucción
tabCon2(dat, "Edad_recod", "Instrucción")
p <- ggplot(dat, aes(x=Instrucción, y=Edad)) +
  geom_boxplot(color = pColores("named")["cAzul0"],
               fill = pColores("named")["cAmaC"])
png(paste0("./Tesis/Latex/Figuras/fg_Edad_Instruccion.png"),
    width = 16, height = 14, units = "cm", res = 300)
p
dev.off()

# Edad y cantón
tabCon2(dat, "Edad_recod", "Cantón")
p <- ggplot(dat, aes(x=Cantón, y=Edad)) +
  geom_boxplot(color = pColores("named")["cAzul0"],
               fill = pColores("named")["cAmaC"])
png(paste0("./Tesis/Latex/Figuras/fg_Edad_Canton.png"),
    width = 16, height = 14, units = "cm", res = 300)
p
dev.off()

# Edad y cerrado
tabCon2(dat, "Edad_recod", "Cerrado")
p <- ggplot(dat, aes(x=Cerrado, y=Edad)) +
  geom_boxplot(color = pColores("named")["cAzul0"],
               fill = pColores("named")["cAmaC"])
png(paste0("./Tesis/Latex/Figuras/fg_Edad_Cerrado.png"),
    width = 16, height = 14, units = "cm", res = 300)
p
dev.off()

```

## Modelo de regresión logística

```

library(ggplot2)
library(xtable)
library(MASS)
library(pROC)

source("MyLib.R")

dat <- read.table("./Datos/emprende.csv",
                 header = TRUE, dec=".", sep = ";")
dat <- subset(dat, select = -c(X,X.1))

mod_sat <- glm(Cerrado ~ Presupuesto + Actividad +
               Instrucción + Edad + Género +
               Tiempo_emprende + Aspectos_emprende +

```

```

        Problemas + Publicidad + Suspendido +
        Personal + Cantón,
data = dat,
family = quasibinomial(link = "logit"))
### Dispersión
sDev <- mod_sat$deviance/mod_sat$df.residual
sDev
### Bondad del ajuste del deviance
pDev <- round(pchisq(mod_sat$deviance, mod_sat$df.residual,
                    lower.tail = FALSE), 4)

pDev
### Hipótesis
res <- summary(mod_sat)
df_res <- as.data.frame(res$coefficients)
#sink(paste0("./Tesis/Latex/Tablas/tb_summary.tex"))
print(xtable(df_res, caption="Análisis de varianza",
             label = "tab_01"),
      caption.placement="top",
      include.rownames=TRUE,
      size="\\fontsize{10pt}{11pt}\\selectfont")
#sink(type="message")
#sink()

### Modelo definitivo
mod_def <- glm(Cerrado ~ Instrucción + Tiempo_emprende +
              Suspendido + Personal + Cantón,
              data = dat,
              family = quasibinomial(link = "logit"))
sDev <- mod_def$deviance/mod_def$df.residual
sDev
### Bondad del ajuste del deviance
pDev <- round(pchisq(mod_def$deviance, mod_def$df.residual,
                    lower.tail = FALSE), 4)

pDev
### Hipótesis
res <- summary(mod_def)
df_res <- as.data.frame(res$coefficients)
#sink(paste0("./Tesis/Latex/Tablas/tb_summary_def.tex"))
print(xtable(df_res, caption="Análisis de varianza para el modelo seleccionado",
             label = "tab_02"),
      caption.placement="top",
      include.rownames=TRUE,
      size="\\fontsize{10pt}{11pt}\\selectfont")
#sink(type="message")
#sink()

df_conf <- confint(mod_def, level = 0.9)

```

```

df_conf[,3]
df_res_exp <- data.frame(Estimación = df_res$Estimate,
                        Exp = exp(df_res$Estimate),
                        Inf90 = exp(df_conf[,1]),
                        Sup90 = exp(df_conf[,2]))
#sink(paste0("./Tesis/Latex/Tablas/tb_coef.tex"))
print(xtable(df_res_exp, caption="Coeficientes e intervalos de confianza al 90\\%",
            label = "tab_coef",
            digits = c(0,2,6,6,6)),
      caption.placement="top",
      include.rownames=TRUE,
      size="\\fontsize{10pt}{11pt}\\selectfont")
#sink(type="message")
#sink()

levels(dat$Cantón)

### ROC
dat$Predic <- predict(mod_def, type = "response")
dat$Pronos <- as.factor(ifelse(dat$Predic<0.2, "No", "Si"))
tab <- xtabs(~Cerrado+Pronos, data = dat)
round(prop.table(tab, margin = 1)*100,2)
res_roc <- roc(Cerrado~Predic, data = dat)
res_roc
p <- ggroc(res_roc, legacy.axes = TRUE,
           color=pColores("named")["cAzul0"]) +
  geom_abline(slope=1,intercept=0, linetype = "dashed",
             color=pColores("named")["cVerde"]) +
  xlab("1 - Especificidad") +
  ylab("Sensibilidad")
p
res_roc
png(paste0("./Tesis/Latex/Figuras/fg_Roc.png"),
    width = 16, height = 14, units = "cm", res = 300)
p
dev.off()

```