Supervivencia de las nuevas empresas

Una aproximación desde el Machine Learning

Daniel Roman

7/17/21

Trabajo de Maestría

Supervivencia de las nuevas empresas.

Una aproximación desde el *Machine Learning*

Daniel Román Ramírez

EAFIT
2021

Supervivencia de las nuevas empresas

Una aproximación desde el *Machine Learning*

Trabajo para optar el título de Magister en Ciencia de Datos

Autor:

Daniel Román Ramírez

Estudiante de Maestría en Ciencia de Datos
Universidad EAFIT
Código: 201920049228

Director:

Edison Valencia Díaz

evalenci@eafit.edu.co
Departamento de Ingeniería de Sistemas

Contenido

Resumen

Emprender es iniciar una búsqueda de generación de valor, a través de la creación o expansión de una actividad económica, por medio de la identificación y explotación de nuevos productos, procesos y mercados.

La generación de emprendimientos depende del ecosistema integrado, que recoge aspectos personales de los individuos, condiciones de mercado, acceso a recursos financieros, políticas públicas por medio de programas y proyectos que favorezcan la formación de negocios.

En Colombia, las estadísticas de los últimos años presentan un crecimiento exponencial de creación de nuevas empresas, sin embargo, más de un 50% de ellas no alcanzan a llegar a los 5 años de vida. Las cifras de fracaso empresarial no han descendido, exponiendo la debilidad de los gobiernos frente al tema de ampliación del desarrollo económico.

Para lograr el aumento en la productividad y la ampliación de la base de la economía, es necesario comprender los desafíos a los que se enfrentan los emprendedores para la supervivencia de la empresa, además, gracias al auge que ha tenido en los últimos años el uso de técnicas del *Machine Learning* y manejo de la información, se aplicarán modelos con base en aprendizaje supervisado y la Regresión Cox, para entender las características importantes del emprendedor y del negocio que podrían afectar la estabilidad en el mercado y con base en ello estimar la probabilidad de supervivencia, en sus primeros años de constituidas. La metodología que se aplicará está basada en modelos de clasificación.

Finalmente, con los resultados del modelo de supervivencia, se espera ser un apoyo útil para los emprendedores aportando información para la toma de decisiones de negocio**.**

Palabras clave: Emprendimiento, Aprendizaje automático, Clasificación, Supervivencia, Mercado, Características, Regresión Cox.

Introducción

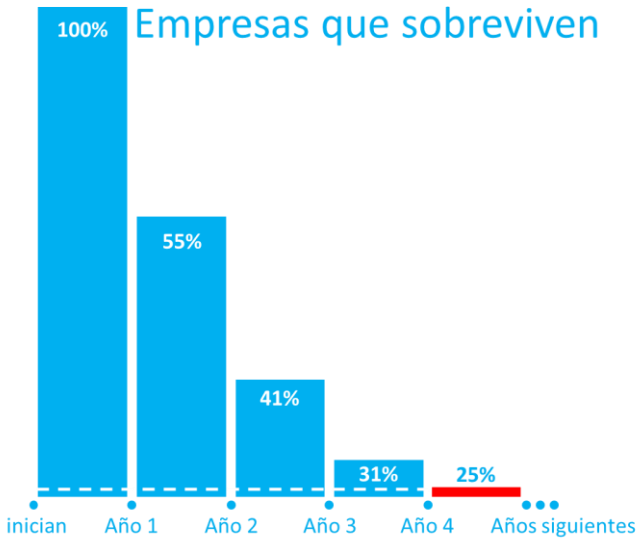
Se entiende por emprendimiento la capacidad de partir de cero en la creación de un proyecto de negocio. A los emprendedores se les denomina hombres o mujeres de negocio. Hoy en día, han ganado importancia estos conceptos por la necesidad de las personas en lograr su independencia y estabilidad económica (Amorós, J. y Guerra, M,2008). Los altos niveles de desempleo y la fragilidad de los empleos existentes, han creado en las personas la necesidad de generar sus propios recursos iniciando sus propios negocios, generando una migración de empleados a empleadores (Confecámaras, 2017).

Los profesionales contemporáneos, consideran, como principal opción de ingreso, el desarrollo de un proyecto de negocio propio y, por tal motivo, los gobiernos han entendido muy bien la importancia del emprendimiento, iniciando programas de apoyo a emprendedores para ayudarles en su propósito de crear su propia unidad productiva (Amorós, J. y Guerra, M,2008).

En Colombia y en los países de la región, los gobiernos disponen de entidades exclusivamente para promover la creación de empresas entre profesionales, y quienes tengan conocimiento específico y suficiente para poder ofertar un producto o servicio lo que genera una confianza a los nuevos emprendedores.

Sin embargo, un estudio realizado por Confecámaras revela que de las empresas que emprenden en Colombia, el primer año subsisten el 55% de ellas, el segundo año el 41% de las que han subsistido y para el tercer año el 31% subsisten de las que han sobrevivido, estimando en el cuarto año un 25% sobreviven de las que han superado tres años.

En la siguiente gráfica se puede observar desde el año 0 (inicio) cuando las empresas se crean (día 0), luego marcando el Año 1 (día 365), que la mitad de ellas han sobrevivido y así progresivamente años tras año, una proporción de las empresas que empezaron en el año 0 han dejado de existir en el mercado.



Según estudios de la CAF (Banco de Desarrollo de América Latina, 2018), países como Colombia, respecto a los países de América Latina, son 2 veces más propensos a la creación de nuevas empresas, pero 6 veces menos capaces de generar compañías con más de 50 empleados. GEM (*Global Entrepreneurship Monitor*, 2018), por su parte, asegura que sólo el 6% de las nuevas empresas resisten más de 3.5 años.

Los investigadores han estudiado este tema a lo largo del espacio y el tiempo, y generalmente lo han hecho a través del análisis cualitativo y considerando cada fase del proceso empresarial por separado (Brixy, Sternberg & Stüber, 2012; Hundt & Sternberg, 2016; van Gelderen, Kautonen, Wincent & Biniari, 2018; entre otros), sin embargo, aún no se encuentra una aplicación real en donde se pueda evidenciar la reducción de la deserción de las pymes en el mercado.

Hoy en día con el uso de técnicas de *Machine Learning*, cambian los paradigmas en las empresas y algunos hábitos que optimizan costos y predicen el comportamiento futuro, entre otros cambios que emplean de forma adecuada los datos. Por tanto, con la ayuda de modelos de clasificación, el proyecto propone calcular la probabilidad de supervivencia de las empresas.

Este trabajo está diseñado de la siguiente manera:

- Clarificación de los conceptos claves del emprendimiento y las evidencias de las estadísticas actuales de las empresas en Colombia, además de estudios externos que se han realizado para comprender las características de supervivencia de las empresas.
- Descripción de la metodología que se llevó a cabo para el desarrollo del proyecto, desde la obtención de los datos hasta la metodología utilizada para alcanzar el objetivo.

- Aplicación de los modelos y presentación de los resultados obtenidos de los modelos y las exploraciones más importantes que se alcanzaron.

Marco conceptual

El concepto de emprendimiento en la sociedad se mide desde hace varios años a nivel global (Amorós y Guerra, 2009), a razón de la vinculación directa con el crecimiento y desarrollo económico de los países (WenneKers y Thurik, 1999).

El emprendimiento se define como la **capacidad de crear un sistema unitario orientado a crear valor y responder a necesidades** (González y Zúñiga, 2011). Una persona emprendedora tiene la habilidad de influir en la conducta de uno o varios actores en el mercado, generalmente para la consecución de finalidades comunes, como: identificando oportunidades, organizando recursos y liderando equipos, es un actor para el desarrollo económico pues pertenecen a la oferta y a la demanda (González y Zúñiga, 2011).

En este contexto, el emprendimiento, ha sido tomado como una opción relevante para cualquier país, sociedad, empresa o individuo. Las estadísticas evidencian que es posible crear oportunidades empresariales cuando se cuentan con los recursos y mecanismos para desarrollar un proyecto emprendedor. El problema se centra en la sostenibilidad y en el impacto que genera el fracaso a los pocos años de constituida (Gutiérrez y Asprilla, 2014).

Un estudio realizado por Confecámaras informa que el 98% de los emprendimientos que mueren en Colombia son microempresas. Según los resultados, factores como el tamaño inicial, el carácter de multiestablecimiento, la orientación al mercado externo y el acceso a créditos a través de garantías mobiliarias, resultan significativos para explicar la dinámica de supervivencia y el riesgo de fracaso de las empresas en Colombia (Confecámaras, 2017).
Amorós y Guerra en GEM de Chile, 2008 (Acs et al,2015,p.38), señala que existe un denominado “círculo vicioso” en donde el emprendimiento contribuye al desarrollo económico y a su vez son las condiciones de desarrollo de los países las que pueden fomentar y fortalecer la creatividad emprendedora y las que generan un desafío importante para la construcción y el diseño de herramientas y de métodos aplicados al emprendimiento que permitan el monitoreo, la evaluación y el desarrollo de la actividad emprendedora, en ambos aspectos con el fin de que la proporción de empresas que fracasen pueda disminuir (Ortiz y Zúñiga,2011).

En consecuencia, diferentes especialistas alrededor del mundo han optado por desarrollar diferentes modelos con aplicaciones matemáticas, económicas y estadísticas con el objetivo de encontrar respuestas a tanta deserción e intentar estimar la supervivencia de las empresas. Estos son algunos ejemplos:

Antecedentes de estudios que estiman la probabilidad de supervivencia de las empresas:

- Leonardo Santana (2017), en su trabajo “Determinantes de la supervivencia de microempresas en Bogotá”, estima mediante modelos de duración la tasa de supervivencia de las microempresas. Inicia calculando las principales razones financieras de rentabilidad, endeudamiento y liquidez, posteriormente aplica modelos de duración, analizando el impacto de cada una de estas variables en la vida de la empresa.
- La aplicación de modelos de duración trata sobre el intervalo de tiempo que transcurre entre el inicio y el fin de un fenómeno. Este trabajo investigativo se remonta en el Reino Unido por Lancaster (1972). Kiefer (1988) realiza una consolidación de los principales desarrollos al respecto, mostrando la función de fallo (Hazard Function) y las distribuciones para los modelos de duración (Weibull, exponencial y log-logística).
- En Estados Unidos tan solo la mitad de las microempresas sobreviven más de cuatro años. Esta tasa es confirmada por Headd(2003), quien mediante modelos de regresión logística, encuentra que el tamaño de la empresa y el capital inicial son factores significativos de éxito.
- Otro modelo de duración aplicado por Lin, Ansell y Andreeva (2012) identificó que cuanto mayor sea el nivel de activos y el acceso a la deuda, mayor es la probabilidad de éxito de las microempresas.
- Una investigación realizada por Glennon y Nigro (2005), confirma que el rechazo de créditos bancarios se debe al número de empleados, es decir, a mayor número de empleados, mayor es el riesgo. Por otro lado, Macas, Goncalves y Serrasqueiro (2013), identificaron que la edad de la empresa afecta el potencial de crecimiento. Agrawal, Chomsisengphet, Liu y Muelnicki (2005) identifican también que ciertas leyes de exenciones de bancarrota en Estados Unidos funcionan como incentivos para declarar la liquidación de las sociedades. Finalmente, Wagner (2013), en su modelo de supervivencia, muestra que la condición exportadora e importadora-exportadora de la empresa afecta positivamente la supervivencia.
- En América Latina, se identificaron algunos artículos que han aplicado modelos de supervivencia. Alcívar y Sainés (2013) analizan determinantes de fallos para empresas en Ecuador, con modelos Cox y Weibull de duración, explicando que uno de los factores significativos de pronóstico de éxito es el sector en el que opera la empresa. Aguilar, Ramírez y Hernández (2011) realizaron un análisis para las empresas en México, y encontraron que existe un fuerte incentivo para que estas empresas sobrevivan manteniéndose en la informalidad. Ortiz (2013) usa modelos Logit para identificar características personales del empresario para predecir el éxito del emprendimiento en República Dominicana. En Colombia, Santana (2014) realiza un modelo Logit con el fin de estimar el costo de capital de una microempresa, asociado a la probabilidad de supervivencia. Parra (2011) realiza una medición mediante modelos Probit con diferentes empresas bogotanas, mostrando que el sector económico y el alto endeudamiento financiero son entre otros, determinantes de la supervivencia de las empresas.

A pesar de los estudios realizados, enfocados al objetivo del proyecto, no se ven su aplicación en la vida real que ayuden a mitigar el fracaso de las empresas.

Metodología

Datos

El desarrollo del proyecto emplea datos en la página oficial de Global Entrepreneurship Monitor. GEM, es una gran organización de investigación internacional que estudia el espíritu empresarial y los factores asociados a este (Bosma, Hill, Ionescu-Somers y Kelley, 2020). La finalidad es unificar las naciones que forman parte del grupo para comprender y observar el emprendimiento como motor del crecimiento económico, comparado entre países (Reynolds, Bosma, Autio y Hunt,2005).

El conjunto de datos es de nivel individual de la Encuesta de población adulta (APS) del Global Entrepreneurship Monitor (GEM), 2016, la cual comprende información detallada sobre la actividad empresarial, las actitudes y las aspiraciones de los encuestados. Esta indagación se aplica en 65 países; el menor número de observaciones en un país es 2,000 y contempla respuestas de emprendedores y no emprendedores.
La base de datos tiene 258 variables y 194,824 observaciones y se divide en 225 variables categóricas y 33 numéricas.

Construcción de la base de datos

Para la construcción de la matriz de datos se filtra la base de datos de GEM por país, seleccionando los registros de Colombia y luego se analizan cada una de las 258 variables que incluye la base.
El conjunto de variables es analizado por la profesora Claudia Patricia Álvarez Barrera quién ha dedicado gran parte de su vida académica estudiando y analizando el entorno emprendedor y conoce a fondo el comportamiento de los micro empresarios, por lo tanto, luego de un análisis detallado, se seleccionan las variables que son importantes para identificar la supervivencia o fracaso del negocio.
Posteriormente se debe completar la base con factores críticos del negocio, como lo son, transacciones por mes, número de empleados, créditos bancarios, sector económico, total ventas mensuales y total ventas anuales. Estas variables fueron validadas con expertos, sin embargo, no se tuvo acceso libre a ellas por ser información reservada de carácter de activo intangible -*Goodwill* de las organizaciones, por lo tanto, se realizan simulaciones aleatorias configuradas con escalas reales y se combinan con las variables ya obtenidas.

Proceso de simulación de datos

Para poder realizar las simulaciones aleatorias, se realizó una investigación sobre las distribuciones que deben tener cada una de las variables numéricas del negocio. Estas indagaciones se estudiaron con personas profesionales que han trabajado en el sector bancario en Colombia (por carácter de datos no se divulgan datos personales).

Toda la aleatoriedad involucrada en el modelo se obtiene a partir de un generador de números aleatorios que produce una sucesión de valores que son realizaciones de una secuencia de variables aleatorias independientes e idénticamente distribuidas (Manuel A Pulido, 2008)

Variable objetivo

La variable objetivo es EXIT_CTD_target, es de carácter binaria e informa si la empresa sobrevivió o no en el mercado después de 3,5 años en el mercado (1: pyme sobrevive, 0: pyme fracasa).
La distribución de esta variable es: 163.322 registros que indican que fracasan y 88.764 registros que indican que sobreviven.

Tabla de variables

A continuación, se describen las variables que pertenecen al conjunto de datos:

Tabla 1: Descripción de las variables.

Nombre variable	Descripción	Valores
ctryalp	*País*	{AE, United Arab Emirates}...
opport	*¿Hay buenas oportunidades para iniciar un negocio en la zona donde vive?*	{0, No, 1, Yes}
suskill	*¿Tiene el conocimiento, la habilidad y la experiencia necesarios para iniciar un nuevo negocio?*	{0, No, 1, Yes}
fearfail	*¿El miedo al fracaso le impediría iniciar un negocio?*	{0, No, 1, Yes}
equalinc	*En mi país, la mayoría de la gente preferiría que todos tuvieran un nivel de vida similar.*	{0, No, 1, Yes}
nbgoodc	*En mi país, la mayoría de la gente considera que iniciar un nuevo negocio es una elección profesional deseable.*	{0, No, 1, Yes}
nbstatus	*En mi país, quienes logran iniciar un nuevo negocio tienen un alto nivel de estatus y respeto.*	{0, No, 1, Yes}
nbmedia	*En mi país, a menudo verá historias en los medios públicos y / o en Internet sobre nuevos negocios exitosos.*	{0, No, 1, Yes}
easystart	*En mi país, es fácil iniciar un negocio.*	{0, No, 1, Yes}
nbsocent	*En mi país, a menudo verá empresas que tienen como objetivo principal resolver problemas sociales.*	{0, No, 1, Yes}
bstart	*¿Está usted, solo o con otros, actualmente tratando de iniciar un nuevo negocio, incluido un trabajo por cuenta propia o la venta de bienes o servicios a otros?*	{0, No, 1, Yes}
bjobst	*¿Está usted, solo o con otras personas, actualmente tratando de iniciar un nuevo negocio o una nueva empresa para su empleador como parte de su trabajo normal?*	{0, No, 1, Yes}
suacts	*Durante los últimos doce meses, ¿ha hecho algo para ayudar a iniciar este nuevo negocio?*	{0, No, 1, Yes}
suown	*¿Será usted personalmente propietario de todo, parte o nada de este negocio?*	{0, No, 1, Yes}
suowners	*¿Cuántas personas, incluido usted, serán propietarios y administrarán este nuevo negocio?*	{0,1,2,3,4….}

suwage	*¿El nuevo negocio ha pagado algún sueldo, salario o pago en especie, incluido el suyo, durante más de tres meses??*	{0, No, 1, Yes}
sucompet	*¿Hay muchas, pocas o ninguna otra empresa que ofrezca los mismos productos o servicios a sus clientes potenciales?*	{0, No, 1, Yes}
sunewtec	*¿Cuánto tiempo han estado disponibles las tecnologías o procedimientos utilizados para este producto o servicio?*	{0, No, 1, Yes}
sunowjob	*Cantidad de personas en la junta directiva*	{0,1,2,3,4… }
suyr5job	*Cantidad de empleados que quieren tener en los próximos 5 años.*	{0,1,2,3,4… }
age	*Edad del emprendedor*	continuo
SUBOANW	*Participa activamente en el esfuerzo de puesta en marcha, propietario, sin salario todavía*	{0, No, 1, Yes}
OMBABYX	*Participación en el esfuerzo de puesta en marcha reclasificada en empresa propia / joven*	{0, No, 1, Yes}
OMESTBX	*Participación en el esfuerzo de puesta en marcha reclasificada en negocio propio / establecido por el hombre*	{0, No, 1, Yes}
BABYBUSM	*Ha administrado una empresa que tiene hasta 42 meses de antigüedad*	{0, No, 1, Yes}
BABYBUSO	*Ha administrado y ha sido propietario de una empresa que tiene hasta 42 meses de antigüedad.*	{0, No, 1, Yes}
ESTBBUSM	*Manages a business that is older than 42 months*	{0, No, 1, Yes}
ESTBBUSO	*Ha Administrado una empresa que tiene más de 42 meses*	{0, No, 1, Yes}
SUBOANWC	*Gestión de propiedad y gestión empresarial reclasificada en creación de empresas*	{0, No, 1, Yes}
BABYBUS1	*Ha clasificado como emprendedor junior?*	{0, No, 1, Yes}
ESTBBUS1	*Ha estado en una reclasificación de emprendedor junior?*	{0, No, 1, Yes}
BUSOWNER	*Ha sido administrador de una empresa que está en funcionamiento?*	{0, No, 1, Yes}
FUTSUPyy	*Espera fortalecer su empresa en los próximos 3 años?*	{0, No, 1, Yes}
DISCENyy	*El negoció no continuó después de 12 meses*	{0, No, 1, Yes}
TEAyy	*Involucrado en la actividad empresarial total en la etapa inicial*	{0, No, 1, Yes}
TEAyyMAL	*Involucrado en la actividad empresarial total en la etapa inicial (Masculino)*	{0, No, 1, Yes}
TEAyyNPM	*TEA: new product market combination*	{0, No, 1, Yes}

SU_JOBNW	*Número de empleados actualmente*	{0,1,2,3,4… }
SU_OWNER	*Número de propietarios actualmente*	{0,1,2,3,4… }
TEAyyHJG	*Esperan tener más de 19 empleados en los próximos 5 años?*	{0, No, 1, Yes}
EXIT_CTD_target	*Empresa sobrevivió después de 3.5 años?*	{0, No, 1, Yes}
EXIT_ENT	*La empresa sobrevivió después de 12 meses?*	{0, No, 1, Yes}
BAFUNDUS	*Fondos actuales de la empresa*	{1000,2000,3000,….}
FRFAILOP	*Miedo al fracaso*	{0, No, 1, Yes}
FUTSUPNO	*Intensiones emprendedoras en el futuro?*	{0, No, 1, Yes}
IPACTNOW_EMP	*Activo como emprendor*	{0, No, 1, Yes}
VARIABLES DEL NEGOCIO		
FECHA_CONST	Fecha de ser constituida la empresa	Fecha
CREDITO_BANC	Total crédito bancario	{100000000,2000000,3000000,… }
CAL_RIESGO_FIN	Calificación del riesgo financiero	(0-100)
SECTOR_ECON	Sector Económico de la EMPRESA	Categórica
NUM_EMPLEADOS	Cantidad de empleados que trabajan en el emprendimiento	{0,1,2,3,4… }
TRX_PROM_MES	Transacciones promedio por mes	{0,1,2,3,4… }
VENTAS_TOTALES_YEAR	Total de ventas en el año	{1000,2000,3000,….}
ANIO_ANTIGUEDAD	Total del tiempo en años de la empresa desde que se constituyó	{1,2,3,4,…..}
MES_PROM_VENTAS	Promedio de ventas al mes	{1000,2000,3000,….}

Método

Para llevar a cabo el estudio, se aplicó el método estándar intersectorial para la minería de datos CRIPS-DM (Shearer,2000), este método es comúnmente usado para aplicaciones de analítica, divide el proceso en diferentes fases de desarrollo (Chapman et al., 2000; Cizallador, 2000):

- Entendimiento de la pregunta de negocio
- Análisis y comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Implementación

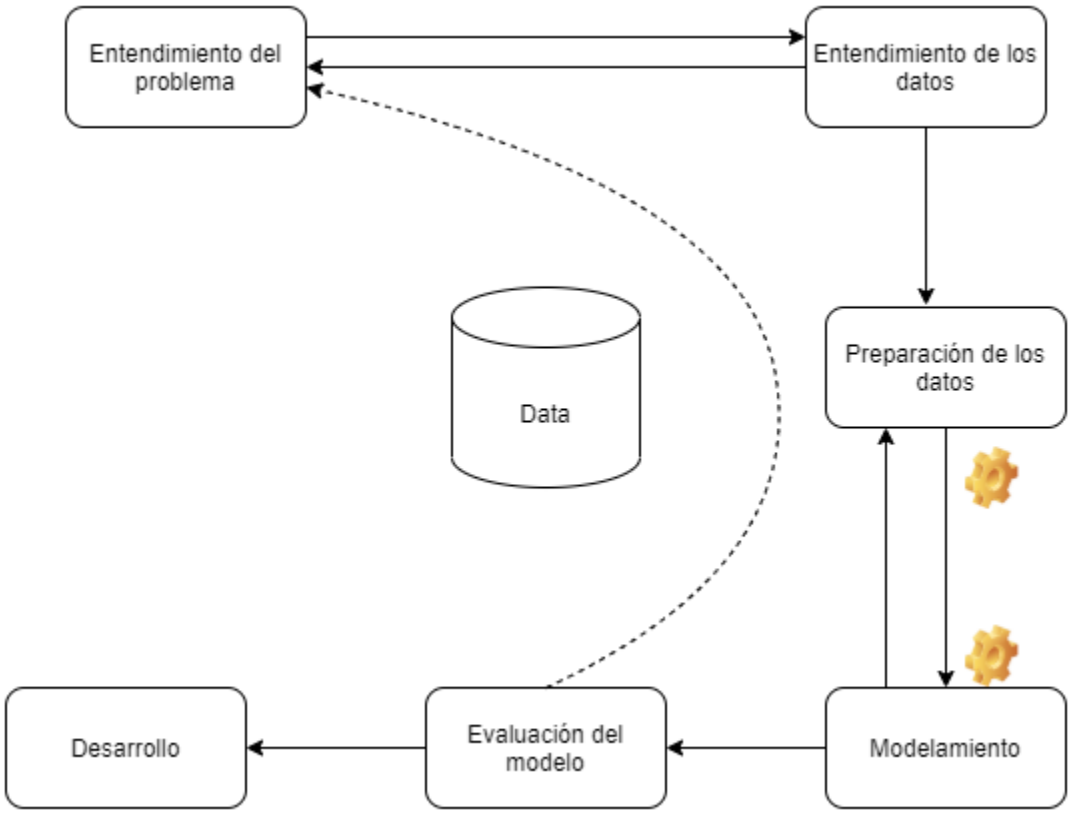


Imagen1: Fases del modelo CRISP-DM, Guía de minería de datos paso a paso,2000, Copyright 1999, 2000 por Pete Chapman, JulianClinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer y Rüdiger Wirth.

Comprensión del problema y datos.

Antes de empezar a conocer y limpiar los datos, se analizaron las estadísticas de los últimos años en Colombia, donde se presentaron un crecimiento exponencial de creación de nuevas empresas. Sin embargo, más de un 50% de ellas no alcanzaron a llegar a los 5 años de vida. Las cifras de fracaso empresarial no han descendido, exponiendo la debilidad de los gobiernos frente al tema de ampliación del desarrollo económico.

Posteriormente, realizando un análisis crítico con todas las variables del conjunto de datos con la experta en emprendimiento, se identificaron variables que no aportan información y que no se relacionan con la supervivencia de la empresa, así que decide eliminarlas, las variables eliminadas son las siguientes:

Tabla 2: Descripción de las variables a eliminar.

Nombre	Etiqueta
CAT_GCR1	Categoría del país
yrsurv	Año de realización de la encuesta
setid	Id
ID	Id 2
weight	Peso del país
WEIGHT_L	Variable GEM
WEIGHT_A	Variable GEM
GEMWORK	Variable GEM
GEMWORK3	Variable GEM
UNEDUC	Variable GEM
knowent	¿Conoce a alguien personalmente que haya iniciado un negocio en los últimos 2 años?
sunewcst	¿Todos, algunos o ninguno de sus clientes potenciales considerarán este producto o servicio nuevo y desconocido?
su_om	¿Es este el mismo negocio al que se refirió en las preguntas anteriores o es un negocio diferente?
occugov	Empleado del gobierno
occunfp	Empleado en el sector voluntario o sin fines de lucro
occuseek	Buscando trabajo
occustu	Estudiante
ipphase1role	¿Y podría decirme si tuvo un papel principal o secundario en esta fase?
ipphase2role	¿Y podría decirme si tuvo un papel principal o secundario en esta fase?
ipinit	¿Fue la actividad nueva más importante en la que estuvo involucrado en los últimos tres años para su empleador principal iniciada por usted mismo, su empleador o uno o más colegas?
ipnewcst	¿Todos, algunos o ninguno de sus clientes potenciales considerarían nuevo y desconocido el producto o servicio desarrollado en este proyecto?

OMTYPE_R	Variable GEM
SUREAS_O	Variable GEM
OMREAS_O	Variable GEM
doublecount	¿Tiene también un teléfono fijo? / ¿También tienes un teléfono móvil?
dtsurv	Fecha de la encuesta
contact	Proveedor de encuestas para registrar cómo se entrevistó al encuestado
callback	Encuesta al proveedor para registrar en qué número de intento de devolución de llamada se contactó con el encuestado (¿1, 2, 3?). Si es entrevistado en el primer contacto, será "0".
TEAISIC4_4D	Código SIC
TEASIC4C	Categoría ISIC

Luego de eliminar las variables que no se utilizarán para los análisis, se adicionan las variables que están relacionadas netamente con el negocio (variables previamente investigadas con la experta en emprendimiento) estas son:

- Transacciones por mes.
- Número de empleados.
- Créditos bancarios.
- Sector económico.
- Total, ventas mensuales.
- Total, ventas anuales.

Esta adición de variables se realiza mediante una simulación aleatoria según la distribución de cada una de ellas y luego se le asigna aleatoriamente a cada registro el cual representa cada emprendedor.
Finalmente se obtiene el conjunto de datos con 252,086 registros y 70 variables (61 variables binarias, 2 categóricas) y 7 numéricas.

Preparación de datos

Con el propósito del estudio de estimar la probabilidad de supervivencia de las Empresas por medio de modelos de clasificación, se preparó la base de datos con la variable objetivo EXIT_CTD_target.

La variable EXIT_CTD_target es de tipo booleana (1: empresa sobrevive, 0: empresa fracasa), es original de la base de datos de GEM e indica si la empresa sobrevivió o no después de 3 años y 6 meses.

El conjunto de datos tiene dos tipos de variables, booleanas y numéricas.
Con el objetivo de identificar qué tipo de variables (psicológicas o de negocio) afectan significativamente la supervivencia de la empresa después de 3 años y 6 meses, se separa

en dos conjuntos de datos: df_empresa_bin (252086 filas y 59 columnas) y df_empresa_num (252,086 dilas y 9 columnas) y así poder ejecutar modelos de clasificación.

Para la preparación de los datos se usa la metodología de análisis exploratorio de datos (EDA) (Behrens, J. T 1997). Utilizando este procedimiento se descubre que no hay evidencias de valores nulos en las dos bases de datos por lo que no se aplican metodologías de imputación de datos o limpieza de valores nulos. Además, se analizan cada una de las características de cada variable binaria y numérica relacionándola con la variable objetivo, con el fin de crear hipótesis que relacionan la supervivencia de la empresa.

Adicionalmente, se identifica que hay desbalance de clases en la variable objetivo pues las clases no están representadas por igual. En este caso se aplica el algoritmo nombrado SMOTE o técnica de sobre muestreo de minorías sintéticas. Este método es un sobre muestreo que funciona creando muestras sintéticas de la clase menor en lugar de crear propias. El algoritmo selecciona dos o más instancias similares y perturba una instancia, un atributo a la vez, en una cantidad aleatoria dentro de la diferencia con las instancias vecinas (NV Chawla, KW Bowyer, LO Hall, W.O. Kegelmwyer, 2002).

Como ya se tienen variables artificiales SMOTE debe trabajar con todo el conjunto de datos y realizar la metodología, no tiene ningún impacto significativo que pueda alterar las predicciones.

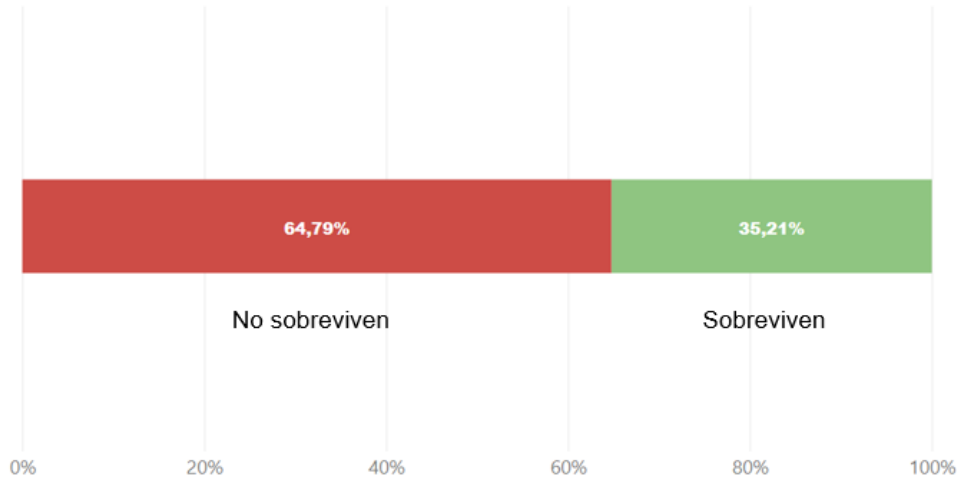


Imagen2: Distribución de clases de la variable objetivo supervivencia de las empresas.

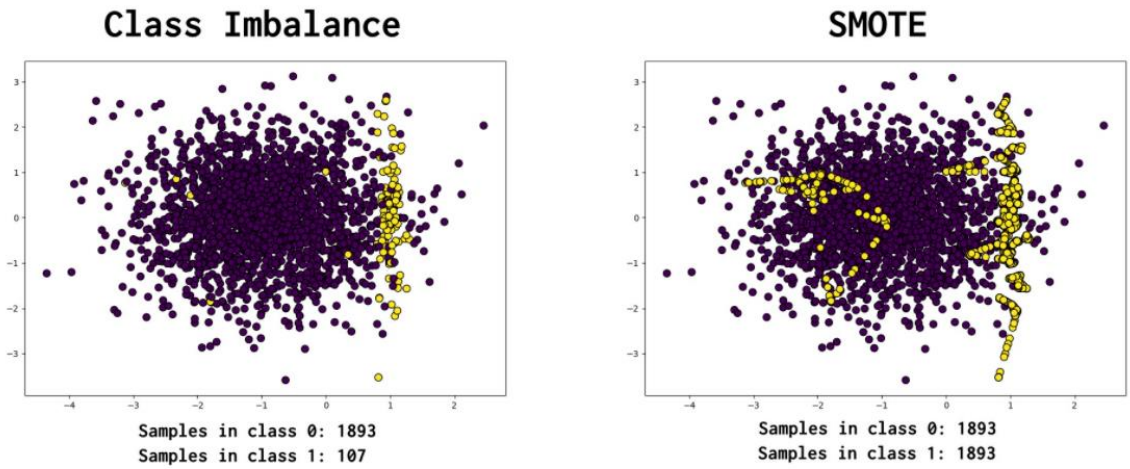


Imagen3: Ejemplo de descripción visual de la metodología de SMOTE (Fernando López, 2021)

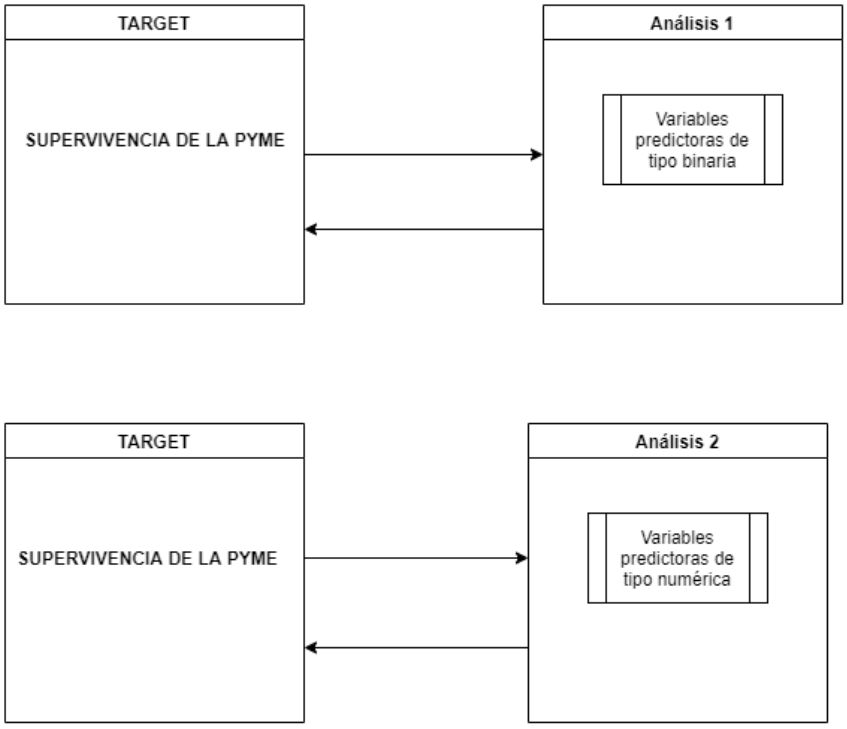


Imagen4: Diagrama de los modelos.

Se obtienen dos conjuntos de datos uno con variables numéricas y otro con variables booleanas. Por cada fase, se realiza un análisis descriptivo de las variables: histogramas y diagramas de caja para las variables numéricas y las fichas cruzadas utilizando las categorías y las clases de la entidad de destino (este análisis se detalla en la sección de Resultados). Se analizan las medidas de resumen agrupadas por las clases de la variable respuesta. Adicionalmente se realizan análisis de correlación entre las variables numéricas y estandarizadas para evitar colinealidad (James, Witten, Hastie & Tibshirani, 2013).

Una vez analizados los dos conjuntos de datos en relación con la variable objetivo, pasamos a la fase de la selección de características y creación de hipótesis según el análisis descriptivo.

La selección de características es el proceso que corresponde a la selección de predictores relevantes para el uso en la construcción de modelos, con el fin de simplificar los modelos y de obtener una mejor interpretación y tiempos de entrenamientos más cortos, para mejorar la generalización mediante la reducción de sobreajuste (Bermingham et al., 2015; James et al., 2013). Debido a que la base de datos de las variables binarias contiene 59 columnas, algunas de ellas son irrelevantes o redundantes (sección descrita en comprensión del problema y datos), y se pueden eliminar sin incurrir en mucha pérdida de la información (Bermingham et al., 2015).

Buscando completar la selección de características, obtenemos la importancia de las variables predictoras en términos del poder predictivo que cada una de ellas tiene y así las relacionamos con la supervivencia de la empresa.

Para ello se utiliza una aplicación de los modelos de clasificación proporcionados por Pycaret (Moez Ali, 2019). Es una librería que permite el modelamiento en *Machine Learning* automático para análisis supervisado y no supervisado, es código libre y abierto para aplicaciones en ciencia de los datos, fue desarrollado originalmente por Moez Ali para Python y R-software. Aplicamos modelos de clasificación, y con el mejor modelo de clasificación (solución detallada en la sección de Resultados) ajustamos los hiper parámetros predeterminados, que proporciona la importancia de las variables predictoras. Posteriormente se analiza la interpretación del mejor modelo, evaluando las métricas y generando las predicciones.

Modelado

Modelos de clasificación

Para identificar el modelo óptimo (Modelo para variables binarias y Modelo para numéricas) se implementaron todos los algoritmos relacionados con modelos de clasificación que ofrece Pycaret, iterando automáticamente y clasificando según la métrica de evaluación, para saber cuál es el mejor de ellos.

Los modelos ejecutados son los siguientes:

- KNN: K Neighbors Classifier
- SVM- linear kernel
- Logistic regression
- Naive Bayes
- Quadratic Discriminat Analysis
- Decision Tree Classifier
- Adaboost Classifier
- Linear Descriminant Analysis
- Ridge Classifier
- Gradient Boosting Classifier
- Extra Trees Classifier
- Random Forest Classifier
- Extreme Gradient Boosting
- Catboost Classifier
- Light Gradient Boosting Machine

La clasificación es una técnica en la que categorizamos los datos en un número determinado de clases, el objetivo principal de problemas de clasificación es identificar la categoría -clase a la que pertenecen los nuevos datos (Rohit Garg, 2018).

Modelo de riesgo proporcional de Cox

El modelo de riesgo proporcional de Cox es un modelo de regresión utilizado generalmente para averiguar la relación entre el tiempo de supervivencia de un sujeto y una o más variables predictoras (Shukla,2020).

El modelo de Cox (riesgo proporcional) es uno de los modelos más populares que combina las covariables y la función de supervivencia. Comienza con el modelado de la función de peligro (Pandey,2019).

$$h(t|X = x) = h_0(t)\exp\,(x^T\beta)$$

β es el vector de coeficientes de cada variable. La función h_0(t) se denomina función de riesgo de línea base.

El modelo de Cox asume que las covariables tienen un efecto de multiplicación lineal sobre la función de riesgo y que el efecto permanece igual a lo largo del tiempo (Pandey,2019)..

De acuerdo a la función anterior se puede derivar la función de riesgo:

$$H(t|x) = \exp\,(x^T\beta)\int_0^t h_o(s)ds = \exp\,(x^T\beta)H_0(t)$$

Por lo que se observa se puede derivar la función de supervivencia a partir de la función de riesgo con la ayuda de la expresión derivada anteriormente y así se deriva la función de supervivencia para cada individuo.

Previo a la modelación es de suma importancia tener claro ciertos conceptos para entender la importancia de estos modelos de clasificación.

Clasificador: Es un algoritmo que asigna los datos de entrada a una categoría específica.

Modelo de clasificación: Un modelo de clasificación intenta sacar predicciones a partir de los valores de entrada dados para el entrenamiento. Predecirá las clases - categorías para los nuevos datos.

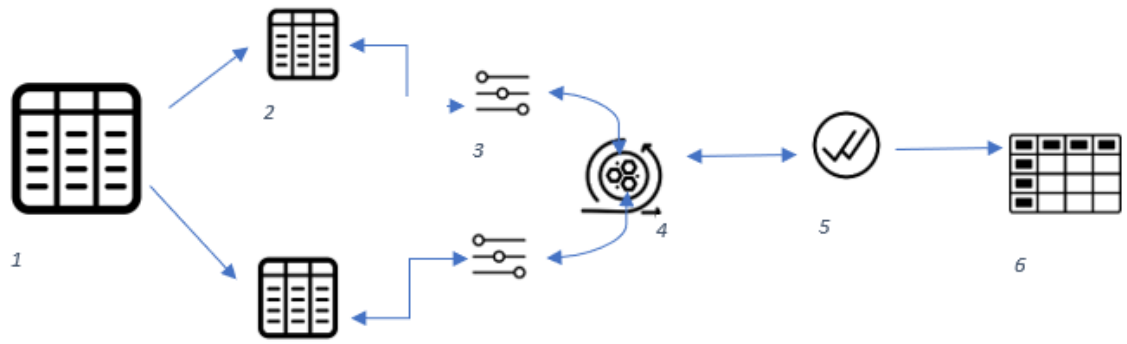
Característica: Una característica es una propiedad individual mensurable de un fenómeno que se observa.

Clasificación binaria: Es una clasificación con dos posibles resultados.

Clasificación multiclase: Es una clasificación con más de dos clases.

Clasificación de etiquetas múltiples: Es una tarea de clasificación en la que cada muestra se asigna a un conjunto de etiquetas de destino (más de una clase).

Pasos involucrados en la construcción del modelo de clasificación.



1. Inicializar el clasificador con la base de datos preparada.
2. Partición de la base del modelo en entrenamiento y prueba
3. Ajuste los hiper parámetros (ajuste del modelo)
4. Entrenamiento del clasificador con el modelo ajustado
5. Evaluación del clasificador
6. Predicciones

Imagen5: Diagrama del flujo del modelo

Diseño del flujo del desarrollo del proyecto

A continuación, se presenta de manera general la metodología empleada en los puntos anteriores:

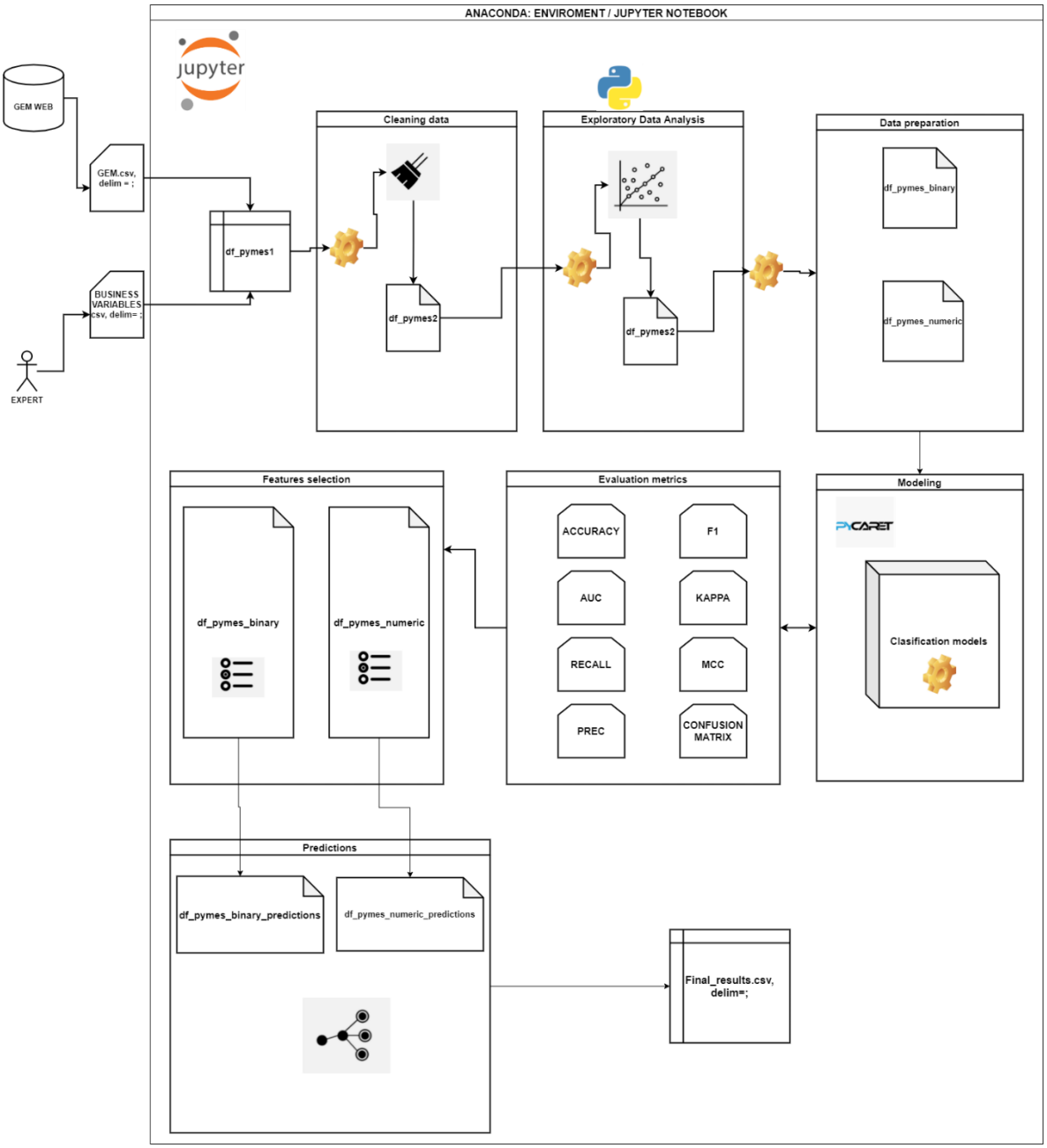


Imagen6: Diagrama del flujo del desarrollo del proyecto

Herramientas que se utilizan

- Excel: Herramienta que se utiliza para la tabulación y consolidación de los datos.
- Anaconda: Herramienta que se utiliza para crear el ambiente con lenguaje Python para el desarrollo técnico del proyecto.
- Jupyter notebook: Herramienta para desplegar el código en Python.
- Pycaret: Librería utilizada en Python para desarrollar modelos de Machine Learning.

Definición de algunos modelos de clasificación.

Regresión logística

Para Hastie, Tibshirani y Friedman (2009), la Regresión Logística es un método lineal para la clasificación. Dado que el predictor toma valores en un conjunto discreto, el espacio de entrada se puede dividir en una colección de regiones etiquetadas según la clasificación y, en este caso, los límites de esas regiones son lineales.

Este modelo es motivado por el deseo de modelar las probabilidades posteriores de las clases K utilizando funciones lineales en x, a la vez que se asegura de que suman a uno y permanecen en [0, 1] al mismo tiempo.

Fórmula de regresión logística:

$$\log\frac{\Pr(G=K-1|X=x)}{\Pr(G=K|X=x)}=\beta_{(K-1)0}+\beta_{(K-1)}^{T}x$$

Imagen7: Fórmula Regresión Logística.

El modelo se define en términos de *K − 1 log-odds* o transformaciones de logit (teniendo en cuenta la restricción que las probabilidades suman a uno). Aunque el modelo utiliza la última clase como denominador en las relaciones de probabilidades, la selección del denominador es aleatoria en el punto de vista de que las estimaciones son equivariantes bajo esta opción. Este modelo se utiliza comúnmente en aplicaciones donde las respuestas son binarias (dos clases) y se producen con bastante frecuencia. Por ejemplo, los pacientes sobreviven o mueren, tienen enfermedades cardíacas sí o no, un préstamo es adjudicado o negado, o una condición está presente o ausente.

Support vector Machine

Hastie, Tibshirani y Friedman (2009) dicen que este modelo produce límites no lineales mediante la construcción de un contorno lineal en una versión grande y transformada del espacio de entidades. La transformación se realiza utilizando una función denominada

Kernel que puede ser lineal o no. Se sugiere para problemas donde un gran número de características están altamente correlacionadas y se puede utilizar para tareas de regresión y clasificación.

Árboles de decisión (Decision Tree)

Decision Tree divide el espacio de características en un conjunto de rectángulos, y luego se ajusta a un modelo simple (como una constante) en cada uno. Son conceptualmente simples pero potentes y se pueden utilizar en problemas de regresión y clasificación.

Intuitivamente, es una estructura similar a un diagrama de flujo con nodos internos, ramas y hojas. Cada nodo interno representa una prueba en un atributo (por ejemplo, si una observación pertenece a un macho o a una hembra), cada rama constituye el resultado de la prueba y cada nodo hoja significa una etiqueta de clase (decisión tomada después de calcular todos los atributos). Las rutas de acceso de raíz a hoja representan reglas de clasificación (Hastie, Tibshirani y Friedman 2009).
En otras palabras, dado un dato de atributos junto con sus clases, un árbol de decisión produce una secuencia de reglas que pueden usarse para clasificar los datos.

Random Forest

El clasificador de bosque aleatorio es un meta estimador que se ajusta a varios árboles de decisión en varias submuestras de conjuntos de datos y usa el promedio para mejorar la precisión predictiva del modelo y controla el sobreajuste. El tamaño de la submuestra es siempre el mismo que el tamaño de la muestra de entrada original, pero las muestras se extraen con reemplazo.

Evaluación

Con el objetivo de evaluar los modelos y evaluar cómo los resultados de los análisis se transforman en otros datos, realizamos una validación *k-fold cross* que consiste en dividir aleatoriamente el conjunto de datos en grupos k, o pliegues, de igual tamaño; el primer pliegue se utiliza como prueba set y los pliegues k-1 restantes se utilizan como tren configurado para ajustarse al modelo. Este método se utiliza principalmente en situaciones en las que el objetivo es predicción y en situaciones cuando es necesario estimar con qué precisión funcionará un modelo (James et al., 2013).

Para este estudio, validamos todas las ejecuciones de todos los modelos en los dos análisis mediante 5 pliegues (5 k), lo que significa que todo el conjunto de datos se divide en cinco conjuntos en cada ejecución. Uno por uno, se selecciona un conjunto como conjunto de pruebas y los cuatro restantes se utilizan como conjunto de secuencias, evaluando todas las combinaciones posibles (cinco iteraciones).

Para cada modelo de cada ejecución, se registran las métricas de rendimiento estadístico para el entrenamiento y las pruebas y las finales se obtienen como el promedio de las puntuaciones de cada iteración (Información expuesta en la sección de Resultados). Se utiliza la precisión como criterio de rendimiento para seleccionar el mejor algoritmo, teniendo en cuenta que en nuestro caso es importante corregir la mayoría de las predicciones.

El *accuracy* (la precisión) se define como una relación entre el total de las muestras clasificadas, correctas o no (Sokolova, Japkowicz & Szpakowicz, 2006; Tharwat, 2020), en otras palabras, es la comparación de las observaciones reales entre las predichas, si la métrica está más cercano a 1 indica que la predicción es confiable, entonces se puede comprender que las variables de negocio ayudan a predecir la supervivencia de las empresas.

La validación cruzada o *cross validation*, se utiliza para evaluar el análisis estadístico cuando el conjunto de datos se ha segmentado en una muestra de entrenamiento y otra de prueba. Esta prueba comprueba si los resultados del análisis son independientes de la partición (Beltran, Mauricio, 2015).

En la siguiente imagen se puede observar cómo es el procedimiento de la validación cruzada.

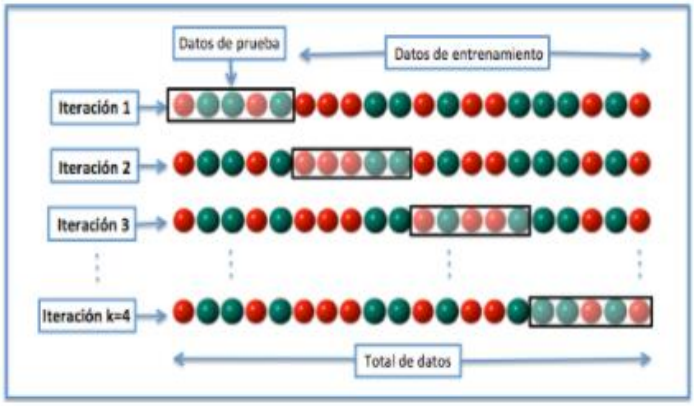


Imagen8: Validación cruzada aleatoria (De Joan.domenech91 - Trabajo propio, CC BY-SA 3.0)

Al final, seleccionamos el mejor modelo con la mejor calificación en las métricas definidas (Información detallada en la sección de Resultados) que puede estimar la supervivencia de la empresa, buscando la menor diferencia entre precisión en formación y pruebas, apelando a una mayor estabilidad del modelo.

Despliegue

Para este trabajo, después de haber llevado a cabo todo el estudio, presentamos un informe de los resultados más importantes y analizamos el tema a la luz de lo que consideramos notable e interesante para el campo de la investigación del emprendimiento.

Resultados

Análisis exploratorio de datos

Luego de tener los conjuntos de datos limpios y estructurados, realizamos un análisis descriptivo y exploratorio de las variables relacionadas con la supervivencia de la empresa.

Edad promedio de los empresarios:

De la base de datos, se identificó que se encuentran empresarios desde una edad mínima de 18 años hasta 70 años. El promedio de edad está entre los 30 y 40 años de edad.

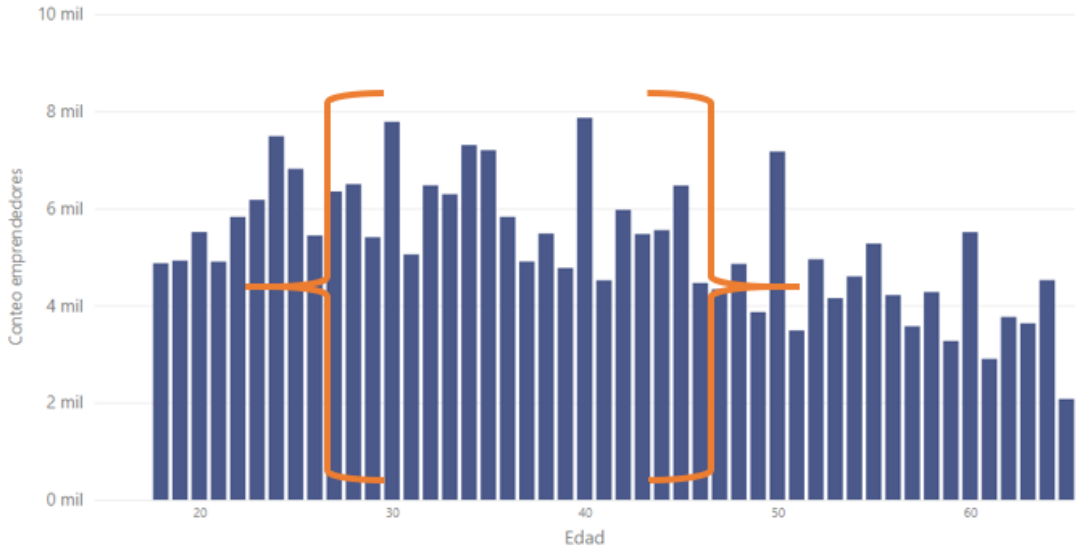


Imagen9: Distribución histograma de la edad promedio de los empresarios

Los resultados entregados por la fase de análisis exploratorio de los datos, resalta las relaciones que tienen las diferentes variables binarias y numéricas con la variable supervivencia de la empresa.

Inicialmente se evidencia la gran proporción de empresas que fracasan rápidamente en el mercado. Solo el 35% de estas sobreviven después de 3.5 años mientras que el 65% restante fracasan.

Y la pregunta es: ¿por qué fracasan tan rápido las empresas en Colombia?, para dar respuesta a esta pregunta, se realizó un análisis exploratorio de cada variable de la base de datos en relación con el éxito de la empresa y estas son las conclusiones al respecto.

- Cuando se pregunta a los empresarios si tuvieron oportunidades de empezar un negocio en el sector donde viven, los que respondieron que sí, el 39.7% sobrevivió mientras que el 60.7% fracasó.
  Los que respondieron que no tuvieron oportunidades de emprender en el sector donde viven solo el 30% sobrevivieron mientras que el 69.4% restante fracasaron.

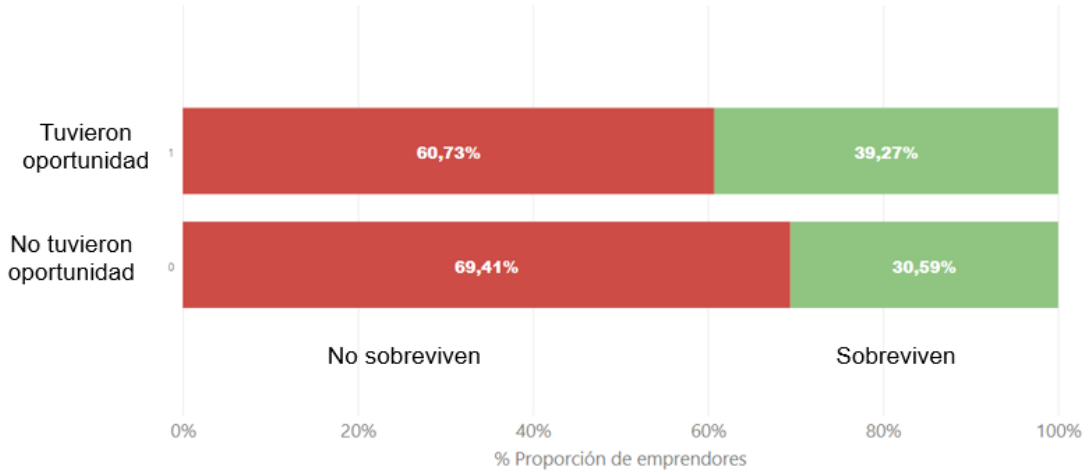


Imagen10: Distribución oportunidades de empezar un negocio en el sector donde viven

- De los que respondieron que sí tienen conocimientos o habilidades previas para emprender, el 44.5% de ellos sobrevivieron al mercado mientras que el 55.4% restante fracasaron. Los que respondieron que no tienen habilidades para emprender o conocimientos previos, y aun así emprendieron, sólo el 21.3% sobrevivieron al mercado mientras que el 76.6% restante fracasaron.

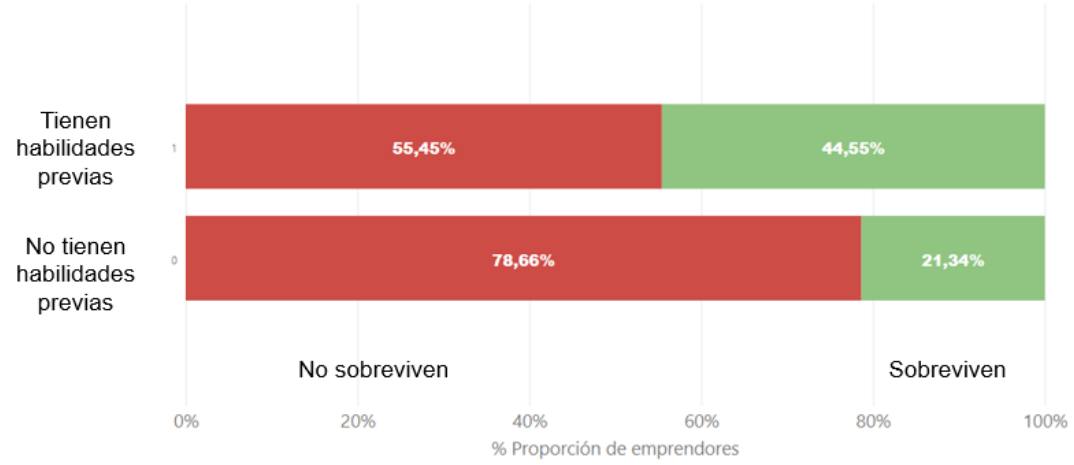


Imagen11: Distribución sobre tener conocimientos o habilidades para emprender

- ¿Sentir miedo o fracaso impide crear un negocio? De los que respondieron que no, el 36.4% de ellos sobrevivieron al mercado mientras que un 63.5% fracasaron. Los que respondieron que sí pero aun así emprendieron, el 33.5% sobrevivieron y el 66.4% fracasaron.

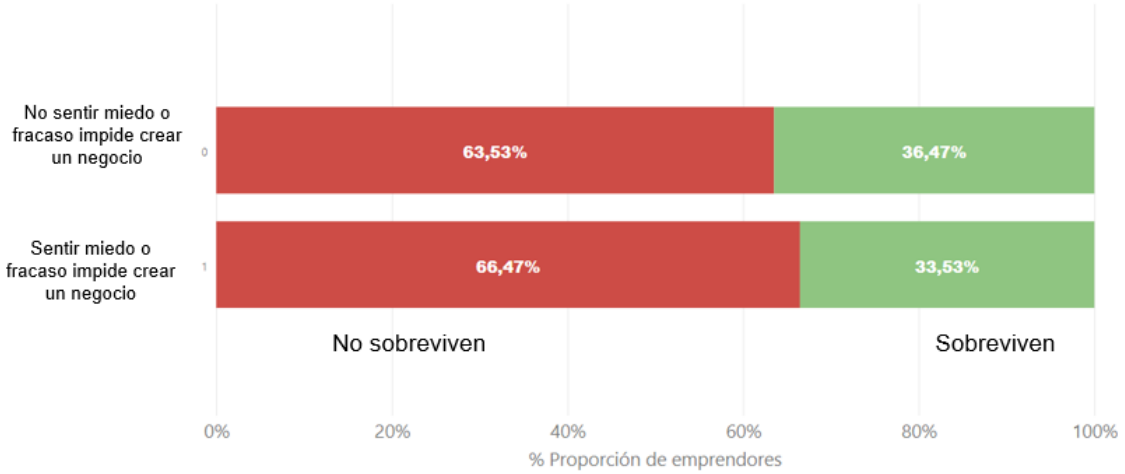


Imagen12: Distribución sentir miedo o fracaso impide crear un negocio.

- ¿Tener vocación de emprendedor es un factor de éxito para sobrevivir? Los que respondieron que sí, el 53.0% sobrevivieron al mercado mientras que el 46.9% fracasaron. Los que respondieron que no pero aun así emprendieron solo el 33.9% sobrevivieron al mercado y el 66.07% fracasaron.

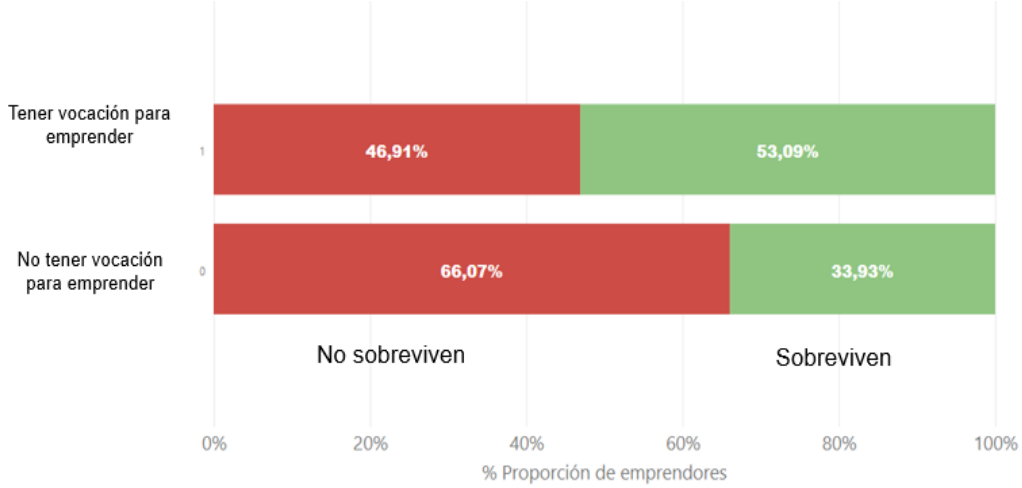


Imagen13: Distribución tener vocación para emprender.

- ¿El país genera facilidad para emprender?, los que respondieron que sí, el 37.3% de ellos sobrevivieron al mercado y el 62.6% fracasaron, los que respondieron que no, el 32.9% sobrevivieron y el 67.0% fracasaron

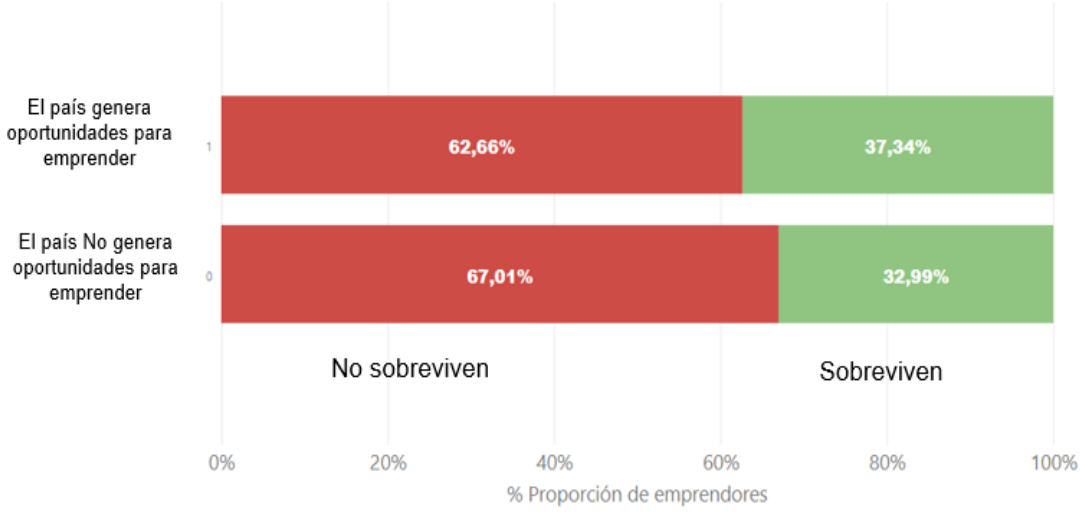


Imagen14: Distribución facilidad de emprender un negocio en el país.

- ¿Registrar los resultados exitosos de las empresas por parte del estado es un factor que afecta a los emprendedores? Los que respondieron que sí, el 37.4% sobrevivieron al mercado. El 62.5% restante fracasaron, los que respondieron que no, el 30.7% sobrevivieron y el 69.5% fracasaron.

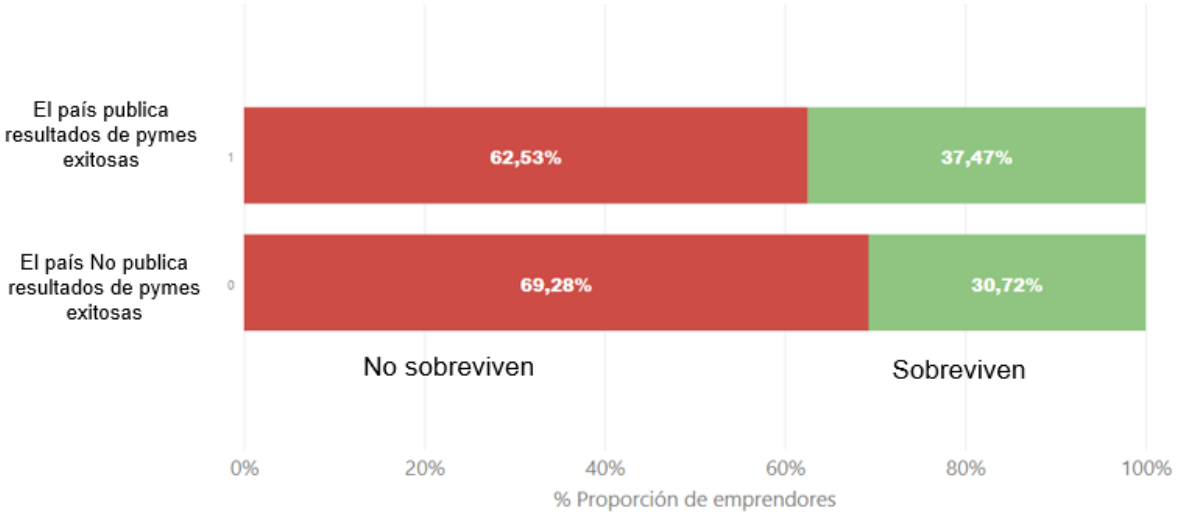


Imagen15: Distribución Publicación de resultados de empresas exitosas

- ¿Iniciar el negocio con un equipo de trabajo está relacionado con el éxito de la empresa? Los que respondieron que sí, el 64% sobrevivieron al mercado y el 35.9%

restante no sobrevivieron. Los que respondieron no, el 30.9% sobrevivieron y el 69.0% fracasaron.

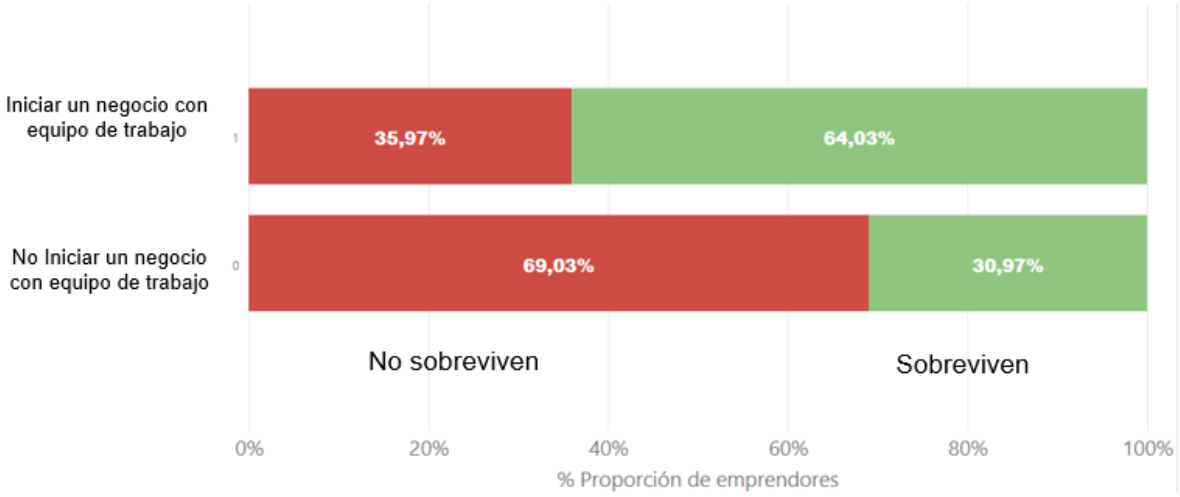


Imagen16: Distribución iniciar un negocio con un equipo de trabajo.

- ¿Tener expectativa de crecimiento de personal en 5 años es un factor de supervivencia? Los que respondieron que sí, el 65.4% sobrevivieron y el 34.5% fracasaron. Los que respondieron que no, el 34.6% sobrevivieron y el 65.3% fracasaron.

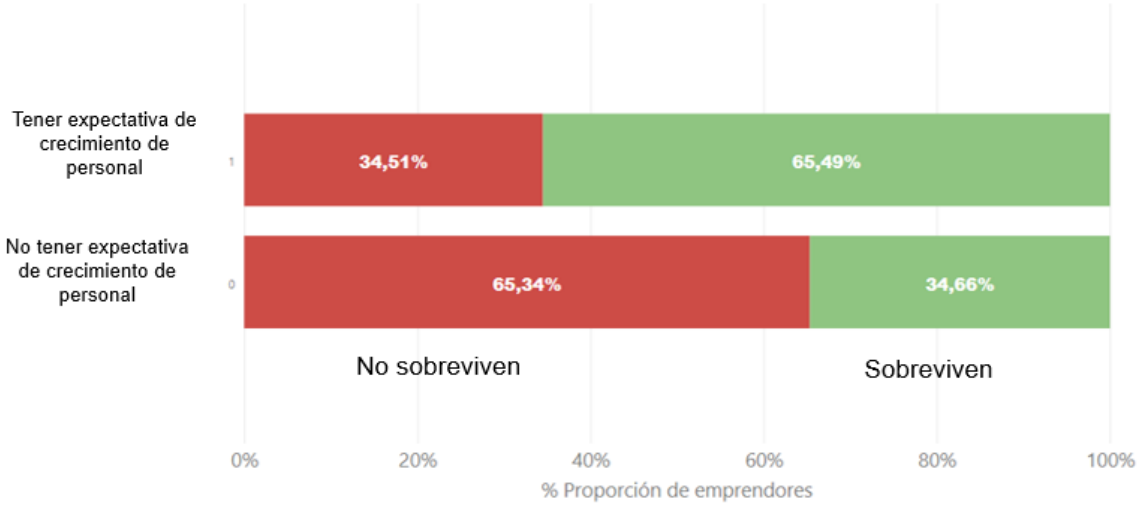


Imagen17: Distribución Expectativa de crecimiento en personal en los próximos 5 años.

Distribución de los sectores comerciales, como se puede evidenciar se observa una mayor participación del sector inmobiliario y turismo en la base de datos.

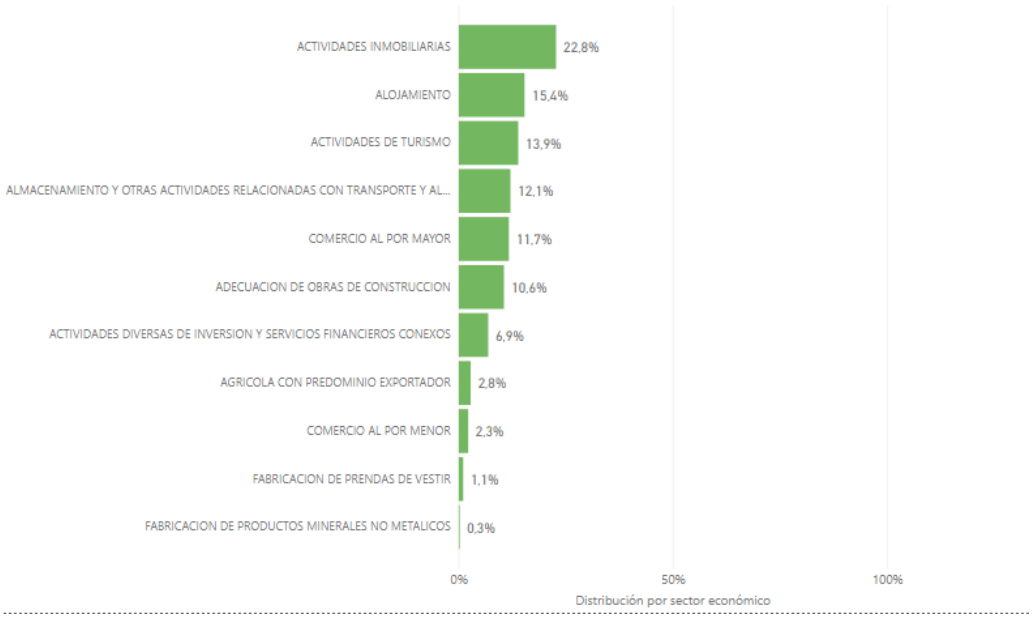


Imagen18: Distribución de los sectores comerciales.

Importancia de las características

En los dos análisis realizados, se obtiene la importancia de la característica en términos de poder predictivo que tiene cada una.

Análisis con las variables binarias usando los modelos de clasificación.

Para este análisis se implementó el modelo que mejor calificación tuvo en el entrenamiento de los datos, ***extra tree classifier (et)*** *(*los resultados comparativos de los otros modelos esta descrita en la sección de Modelos). Esta metodología es una técnica de aprendizaje por conjuntos que agrega los resultados de varios árboles de decisión no correlacionados recopilados en un "bosque" para generar su resultado de clasificación. En pocas palabras, es muy similar a un clasificador de bosque aleatorio y solo se diferencia de él en la forma de construcción de los árboles de decisión en el bosque (AlindGupta, 2020)

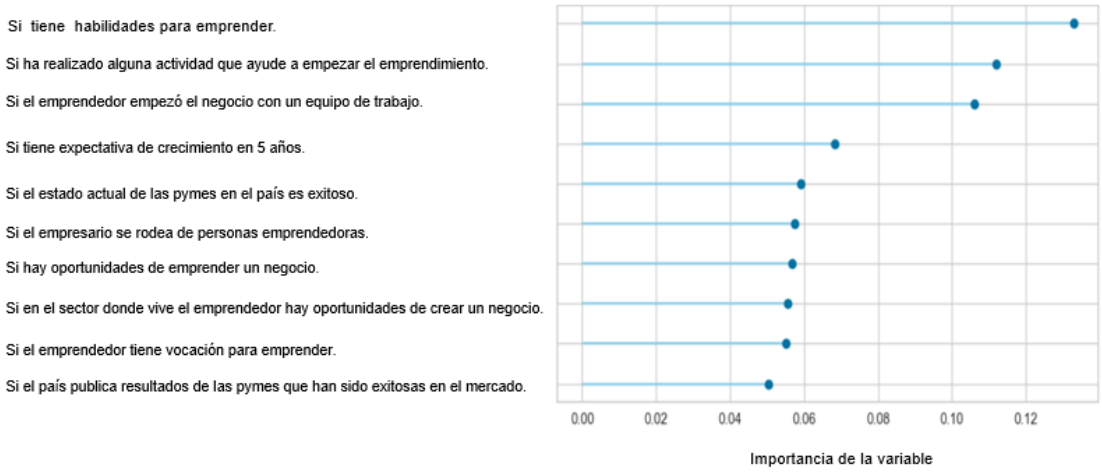


Imagen19: Características importantes de tipo binaria para la predicción de supervivencia de las Empresas.

Características de las variables binarias.

De la gráfica obtenemos lo siguiente:

- Si tiene habilidades para emprender (Suskill = 1):
  Tener habilidades o conocimientos previos para emprender es un factor que influye significativamente en una alta probabilidad de supervivencia.
- Si ha realizado alguna actividad que ayude a empezar el emprendimiento (Suacts =0) :
  Esta variable se refiere si en los últimos 12 meses ha realizado alguna actividad que ayude a empezar el emprendimiento. En este caso no haber realizado alguna actividad que influya en el inicio del emprendimiento es un factor importante para la supervivencia de la empresa.
- Si el emprendedor empezó el negocio con un equipo de trabajo (Bstart = 0):
  Esta variable se refiere a que si el emprendedor empezó el negocio con un equipo de trabajo. En este caso no tener un equipo de trabajo para el inicio del proyecto emprendedor influye de manera negativa en la probabilidad de supervivencia.
- Si tiene expectativa de crecimiento en 5 años (TEAayyHJG = 0):
  Esta variable se refiere a la expectativa de crecimiento de empleados en 5 años. Para esta situación, no tener expectativas de crecimiento del número de empleados influye en la probabilidad de supervivencia.
- Si el estado actual de las Empresas en el país es exitoso (Nbstatus =1):
  Se refiere al estado actual en el país de las empresas nuevas en el mercado. Los empresarios si les importa el estado actual de las empresas en el país es un factor positivo para aumentar la probabilidad de supervivencia.
- Si el empresario se rodea de personas emprendedoras (Equalinc = 0):
  Se refiere al estilo de vida de las personas que rodean al empresario, es decir si se rodea de personas exitosas. En este caso cuando no se rodea del mismo estilo de mentalidad o vida de las personas es un factor significativo en la probabilidad de supervivencia.

- Si hay oportunidades de emprender un negocio (Easystart =1):
  Se refiere a la facilidad que da el país para emprender un negocio. Cuando se presentan oportunidades y facilitan el libre emprendimiento aumentan la probabilidad de supervivencia.
- Si en el sector donde vive hay oportunidades de emprender (Opport =1):
  Se refiere a las oportunidades que se dan para emprender un negocio en el sector donde vive el emprendedor. Cuando hay estas oportunidades, la probabilidad se ve afectada positivamente.
- Si el emprendedor tiene vocación para emprender (Nbgoodc = 0):
  Se refiere a que si la persona tiene vocación de emprendedor. En este caso no tener vocación de emprendedor es una causal de que la probabilidad de supervivencia sea afectada negativamente.
- Si el país publica resultados de las Empresas que han sido exitosas en el mercado (Nbmedia=0):
  Se refiere cuando el país publica los resultados de las empresas exitosas en el mercado. El emprendedor siempre está pendiente de las estadísticas y este es un factor importante para que la probablidad de supervivencia crezca.

Análisis con las variables numéricas usando los modelos de clasificación.

Según el mejor modelo de clasificación que es ***Catboost Classifier*** **(**La comparación de todos los modelos esta descrita en la sección de Modelos), identificamos las variables que son importantes para la predicción de supervivencia de las Empresas.
Catboost es un algoritmo automático de código abierto de Yandex (Sunil Ray,2017).

Cat viene de la palabra "***Category*****"** y boost de "***Boosting*****".** Boost nace del algoritmo de aprendizaje automático que impulsa el gradiente. El aumento de gradiente es un algoritmo muy poderoso de aprendizaje automático que se aplica ampliamente a múltiples tipos de desafíos comerciales, como detección de fraudes, elementos de recomendación, pronósticos y también funciona bien. También puede devolver muy buenos resultados con relativamente menos datos, a diferencia de los modelos DL que necesitan aprender de una gran cantidad de datos (Sunil Ray,2017).

Para este modelo se tienen las siguientes variables importantes.

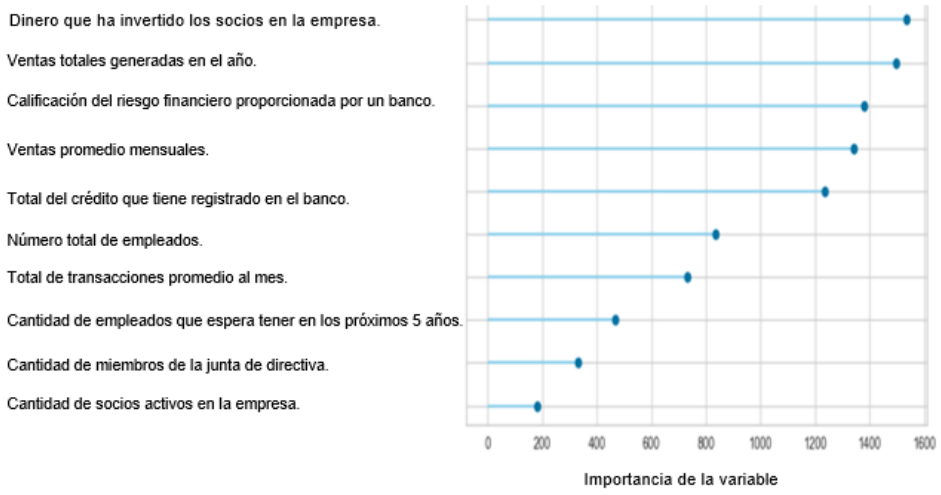


Imagen 20: Características importantes de tipo numérica para la predicción de supervivencia de las Empresas.

Características numéricas

De la gráfica obtenemos lo siguiente:

- Dinero que ha invertido los socios en la empresa (BAFUNDS):
  Se refiere a los fondos que han invertido los socios de la empresa.
- Ventas totales generadas en el año (VENTAS_TOTALES_YEAR):
  Se refiere al total de ventas de la empresa generadas en todo el año.
- Calificación del riesgo financiero proporcionado por un banco (CAL_RIESGO_FIN):
  Esta variable se refiere a la calificación del riesgo financiero que hace alguna entidad bancaria le asigna alguna persona natural o jurídica según su historial bancario.
- Ventas promedio mensuales (MES_PROM_VENTAS):
  Se refiere al promedio de ventas de la empresa generadas en cada mes.
- Total de crédito que tiene registrado en el banco (CREDITO_BANC):
  Está relacionada al total del crédito bancario emitido por la entidad financiera a la persona natural o jurídica.
- Número total de empleados (NUM_EMPLEADOS):
  Se refiere al total de empleados que tiene cada empresa.
- Total de transacciones promedio al mes (TRX_PROM_MES):
  Se refiere a la cantidad de transacciones en ventas realizadas mes a mes.
- Cantidad de empleados que espera tener en los próximos 5 años (Suyr5job):
  Esta variable indica cuántos empleados esperan crecer en los próximos 5 años.
- Cantidad de los miembros de la junta directiva (Sunowjob):
  Esta variable indica cuántas personas están relacionadas con la junta directiva.
- Cantidad de socios activos en la empresa (Suowners):
  Esta variable indica cuántas personas son socias activas en la empresa.

- Inversión en nueva tecnología (sunewtec):
  Esta variable indica la inversión en tecnología para la empresa.

Análisis con las variables numéricas usando el modelo Cox.

El modelo de regresión Cox utiliza la métrica valor-p para identificar las variables que son significativas (Wasserstein R, 2016), partiendo de la hipótesis nula de que si el valor-p es menor que 0.05 (Valor-p<0.05) significa que tiene una relación fuerte sobre la supervivencia de la empresa.

En la tabla 3 se puede observar todas las variables numéricas y se sobresalta las variables que tienen un valor-p menor que 0.05.

Tabla 3: Variable numéricas usadas por el Modelo Cox.

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
CREDITO_BANC	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-1.00	0.32	1.65
CAL_RIESGO_FIN	0.01	1.01	0.00	0.01	0.01	1.01	1.01	77.51	<0.005	inf
NUM_EMPLEADOS	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-0.03	0.98	0.03
TRX_PROM_MES	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-0.57	0.57	0.82
VENTAS_TOTALES_YEAR	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.49	0.62	0.69
MES_PROM_VENTAS	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-0.54	0.59	0.76
BAFUNDUS	0.00	1.00	0.00	0.00	0.00	1.00	1.00	26.45	<0.005	509.82
sunewtec	0.79	2.21	0.01	0.78	0.81	2.18	2.25	99.05	<0.005	inf
sunowjob	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.13	0.89	0.16
suowners	0.00	1.00	0.00	0.00	0.00	1.00	1.00	2.33	0.02	5.66
suyr5job	0.00	1.00	0.00	0.00	0.00	1.00	1.00	7.32	<0.005	41.90

Análisis con las variables binarias usando el modelo Cox.

De acuerdo al modelo de regresión Cox, las siguientes variables binarias son significativas partiendo de la hipótesis nula de que si el Valor -p es menor que 0.05 (Valor-p<0.05) significa que tiene una relación fuerte sobre la supervivencia de la empresa.

Tabla 4: Variables binarias usadas por el Modelo Cox.

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
opport	0.01	1.01	0.01	-0.01	0.02	0.99	1.02	0.94	0.35	1.53
suskill	0.54	1.72	0.01	0.52	0.56	1.69	1.75	64.85	<0.005	inf
fearfail	-0.01	0.99	0.01	-0.02	0.00	0.98	1.00	-1.43	0.15	2.71
equalinc	-0.06	0.94	0.01	-0.07	-0.04	0.93	0.96	-7.77	<0.005	46.90
nbgoodc	0.06	1.06	0.01	0.04	0.07	1.05	1.08	7.68	<0.005	45.89
nbstatus	-0.04	0.96	0.01	-0.05	-0.02	0.95	0.98	-5.02	<0.005	20.91
nbmedia	0.10	1.11	0.01	0.09	0.12	1.09	1.12	13.18	<0.005	129.43
easystart	-0.08	0.92	0.01	-0.09	-0.06	0.91	0.94	-10.78	<0.005	87.55
nbsocent	0.14	1.16	0.01	0.13	0.16	1.14	1.17	19.55	<0.005	280.43
bstart	0.30	1.35	0.01	0.27	0.32	1.32	1.38	26.37	<0.005	506.77
bjobst	0.18	1.20	0.01	0.16	0.20	1.18	1.23	18.45	<0.005	250.02
suacts	0.43	1.54	0.02	0.39	0.47	1.48	1.60	21.30	<0.005	332.08
suown	-0.27	0.77	0.02	-0.31	-0.22	0.73	0.80	-11.58	<0.005	100.59
suwage	0.29	1.34	0.01	0.27	0.32	1.30	1.37	22.60	<0.005	373.26
sucompet	0.03	1.03	0.01	0.00	0.06	1.00	1.06	2.24	0.03	5.32
TEAyyHJG	-0.35	0.70	0.01	-0.37	-0.34	0.69	0.71	-52.34	<0.005	inf

De la tabla 4 se puede observar que la mayoría de variables son significativas para el modelo y que están relacionadas con la supervivencia de la empresa, solamente se excluyeron fearfail que indica si la persona emprendedora tiene miedo al fracaso o no, y sucompet que se refiere que si la competencia vende los mismos productos.

Modelos

En la siguiente sección se describirán los modelos de clasificación ejecutados para los dos conjuntos de variables en relación a la supervivencia de las empresas.

Análisis de las métricas usadas para los modelos de clasificación con las variables binarias.

En la siguiente tabla, se observan los resultados de los modelos que se entrenaron con el conjunto de variables tipo binaria.

Tabla 5: Modelos de clasificación usados para predecir la supervivencia de las empresas

Model	Accuracy	AUC	Recall	Prec	F1	Kappa	MCC	TT(Sec)
ExtraTrees Classifier	0.7245	0.7855	0.6717	0.5958	0.6320	0.4131	0.4150	30.7
Random Forest Classifier	0.7232	0.7873	0.6750	0.5942	0.6320	0.4116	0.4138	39.5
Decision Tree Classifier	0.7220	0.7832	0.6727	0.5929	0.6320	0.4090	0.4112	18.0
K Neighbors Classifier	0.7105	0.7227	0.5219	0.6030	0.5592	0.3022	0.3476	72.3
Catboost Classifier	0.7050	0.7667	0.6551	0.5707	0.6100	0.3745	0.3769	57.3
Naive Bayes	0.6998	0.7145	0.3868	0.6179	0.4557	0.2807	0.2986	16.6
Extreme Gradient Boosting	0.6993	0.7584	0.6470	0.5636	0.6024	0.2990	0.3038	21.93
Quadratic Discriminant Analysis	0.6957	0.7204	0.4308	0.5937	0.4993	0.3472	0.3504	16.0
SVM-Linear Kernel	0.6919	0.000	0.4856	0.5765	0.5244	0.3192	0.3202	17.93
Light Gradient Boosting Machine	0.6898	0.7495	0.6505	0.5504	0.5962	0.3187	0.3197	17.94
Rigde Classifier	0.6818	0.000	0.6014	0.5435	0.5710	0.3192	0.3202	21.8
Linear Discriminant Analysis	0.6815	0.7219	0.6014	0.5436	0.5708	0.3192	03202	17.9
Ada Boost Classifier	0.6812	0.7222	0.6013	0.5436	0.5708	0.3187	0.3197	20.6
Logistic Regression	0.6763	0.7223	0.6028	0.5432	0.5711	0.3186	0.3197	17.9
Gradient Boosting Classifier	0.6763	0.7269	0.6361	0.5426	05805	0.3202	0.3235	30.9

Tabla2: Modelos de clasificación entrenados con las variables binarias.

De los resultados obtenidos, se considera que el mejor modelo en todas las métricas es el **Extra Trees Classifier**. Los modelos evaluados con equilibrio de clases no producen resultados sobresalientes, es decir, se esperaban resultados por encima de 0,80 en *accuracy* (precisión). Lo que significa que el modelo tiene una confiabilidad en las predicciones de clasificación si la empresa sobrevive o no del 72% siendo este valor no tan bueno como lo podría ser un valor por encima de 80%, referenciando que, entre más cercano esté la calificación del *accuracy* a 100% es más confiable el modelo.

Ajuste de hiper parámetros para el Extra Trees Classifier en Pycaret

Para realizar el ajuste de hiperparámetros se usó la opción de Tune-sklearn de pycaret la cuál es un reemplazo directo del módulo de selección de modelos de scikit-learn. tune-sklearn proporciona una API unificada basada en scikit-learn que le brinda acceso a varios algoritmos y bibliotecas de optimización de última generación, incluidos Optuna y scikit-Optimize. Esta API unificada le permite alternar entre muchas bibliotecas de optimización de hiperparámetros diferentes con un solo parámetro (Baum, 2021).

tune-sklearn funciona con Ray Tune, una biblioteca de Python para la ejecución de experimentos y el ajuste de hiperparámetros a cualquier escala. Esto significa que puede escalar su ajuste en varias máquinas sin cambiar su código (Baum, 2021).

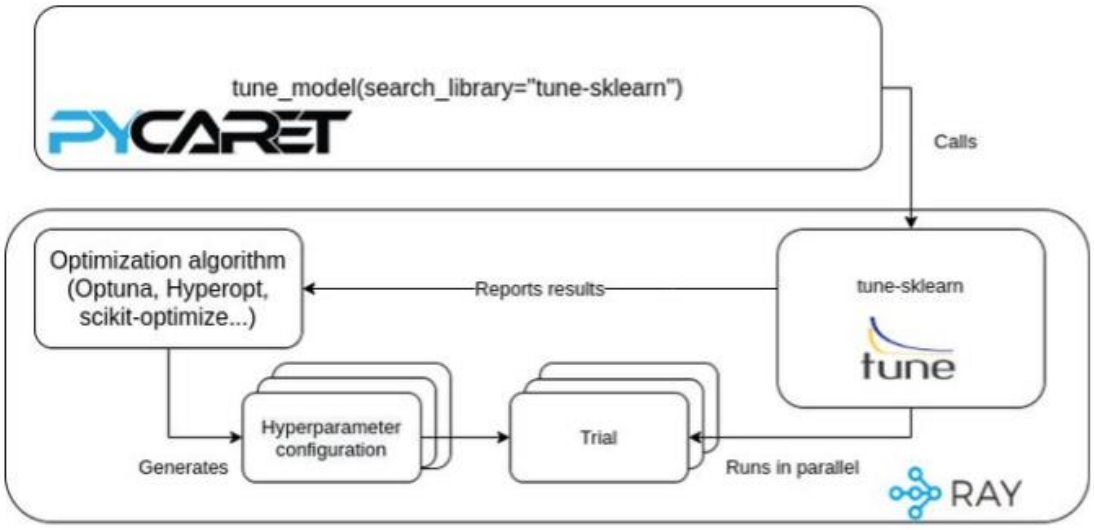


Fig.: Funcionamiento de Pycaret (img tomada de zephyrnet.com/bayesian-hyperparameter-optimization-with-tune-sklearn-in-pycaret

Para lograr mejorar el modelo seleccionado se deben ajustar los hiper parámetros. En la siguiente tabla observamos los resultados obtenidos.

Tabla6: Ajuste del modelo para mejorar la confiabilidad de las predicciones

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7010	0.6182	0.3383	0.6434	0.4435	0.2651	0.2907
1	0.6981	0.6153	0.3356	0.6347	0.4390	0.2584	0.2830
2	0.7016	0.6198	0.3430	0.6430	0.4474	0.2681	0.2929
3	0.6980	0.6153	0.3359	0.6343	0.4392	0.2584	0.2829
4	0.7026	0.6207	0.3439	0.6460	0.4489	0.2703	0.2955
5	0.6999	0.6197	0.3487	0.6344	0.4501	0.2669	0.2895
6	0.6994	0.6170	0.3386	0.6378	0.4423	0.2621	0.2867
7	0.6996	0.6164	0.3347	0.6408	0.4397	0.2612	0.2869
8	0.7018	0.6183	0.3360	0.6476	0.4425	0.2658	0.2925
9	0.7023	0.6204	0.3433	0.6454	0.4482	0.2695	0.2947
Mean	0.7004	0.6181	0.3398	0.6407	0.4441	0.2646	0.2895
SD	0.0016	0.0019	0.0044	0.0049	0.0040	0.0041	0.0043

Creación del modelo Extra Trees Classifier para las variables binarias

Finalmente se crea el modelo entrenando con 10 iteraciones con el ajuste de los hiper parámetros y obtenemos un *accuracy* de 0,72 en promedio, lo que nos indica que el modelo tiene una confiabilidad del 72% en las predicciones de clasificación de la supervivencia de la empresa. Como podemos observar no hubo una mejora significativa en el accuracy del modelo por lo que se sostiene en un poder predictivo del 72%, esto quiere decir que el modelo puede indicar que la empresa va a sobrevivir en el mercado con una confiabilidad del 72% de acuerdo a ciertas variables que están relacionadas con la supervivencia.

Tabla7: Calificación del modelo seleccionado

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7219	0.7873	0.6741	0.5923	0.6305	0.4090	0.4112
1	0.7163	0.7766	0.6614	0.5861	0.6214	0.3959	0.3978
2	0.7246	0.7860	0.6710	0.5969	0.6318	0.4130	0.4148
3	0.7215	0.7893	0.6781	0.5910	0.6316	0.4094	0.4119
4	0.7258	0.7871	0.6693	0.5990	0.6322	0.4147	0.4163
5	0.7287	0.7917	0.6790	0.6018	0.6380	0.4224	0.4243
6	0.7271	0.7881	0.6814	0.5989	0.6375	0.4201	0.4223
7	0.7330	0.7919	0.6683	0.6104	0.6380	0.4272	0.4283
8	0.7267	0.7870	0.6695	0.6004	0.6330	0.4163	0.4179
9	0.7202	0.7809	0.6646	0.5914	0.6259	0.4036	0.4053
Mean	0.7246	0.7866	0.6717	0.5968	0.6320	0.4131	0.4150
SD	0.0045	0.0044	0.0061	0.0066	0.0051	0.0087	0.0086

Modelo Cox con las variables binarias

En la siguiente tabla se exponen los resultados obtenidos del modelo Cox para las variables binarias.

Tabla 8: Resultados del modelo Cox para las variables binarias.

model	lifelines.CoxPHFitter
duration col	'ANIO_ANTIGUEDAD'
event col	'EXIT_CTD_target'
baseline estimation	breslow
number of observations	252086
number of events observed	88764
partial log-likelihood	-1024908.97
time fit was run	2021-10-29 23:23:32 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
opport	0.01	1.01	0.01	-0.01	0.02	0.99	1.02	0.94	0.35	1.53
suskill	0.54	1.72	0.01	0.52	0.56	1.69	1.75	64.85	<0.005	inf
fearfail	-0.01	0.99	0.01	-0.02	0.00	0.98	1.00	-1.43	0.15	2.71
equalinc	-0.06	0.94	0.01	-0.07	-0.04	0.93	0.96	-7.77	<0.005	46.90
nbgoodc	0.06	1.06	0.01	0.04	0.07	1.05	1.08	7.68	<0.005	45.89
nbstatus	-0.04	0.96	0.01	-0.05	-0.02	0.95	0.98	-5.02	<0.005	20.91
nbmedia	0.10	1.11	0.01	0.09	0.12	1.09	1.12	13.18	<0.005	129.43
easystart	-0.08	0.92	0.01	-0.09	-0.06	0.91	0.94	-10.78	<0.005	87.55
nbsocent	0.14	1.16	0.01	0.13	0.16	1.14	1.17	19.55	<0.005	280.43
bstart	0.30	1.35	0.01	0.27	0.32	1.32	1.38	26.37	<0.005	506.77
bjobst	0.18	1.20	0.01	0.16	0.20	1.18	1.23	18.45	<0.005	250.02
suacts	0.43	1.54	0.02	0.39	0.47	1.48	1.60	21.30	<0.005	332.08
suown	-0.27	0.77	0.02	-0.31	-0.22	0.73	0.80	-11.58	<0.005	100.59
suwage	0.29	1.34	0.01	0.27	0.32	1.30	1.37	22.60	<0.005	373.26
sucompet	0.03	1.03	0.01	0.00	0.06	1.00	1.06	2.24	0.03	5.32
TEAyyHJG	-0.35	0.70	0.01	-0.37	-0.34	0.69	0.71	-52.34	<0.005	inf

Concordance	0.66
Partial AIC	2049849.94
log-likelihood ratio test	25669.90 on 16 df
-log2(p) of ll-ratio test	inf

De la tabla 8 se puede observar que el modelo tiene una confiabilidad del 0.66, lo que indica que el modelo no es tan bueno para las predicciones, sin embargo el modelo identifica las variables significativas según el valor p y además expone los coeficientes que deberían de tener estas variables para las predicciones de supervivencia de cada una de las empresas.

Modelo Cox con las variables numéricas

En esta sección analizaremos los resultados obtenidos con el modelo de Regresión Cox.
A continuación se observa el resumen de cada una de las variables numéricas que están en relación con la variable tiempo "ANIO_ANTIGUEDAD" que significa el tiempo de las empresas desde que fueron constituidas y con la supervivencia de las empresas (EXIT_CTD_target).

Tabla 9: Modelo Cox para las variables numéricas.

model	lifelines.CoxPHFitter
duration col	'ANIO_ANTIGUEDAD'
event col	'EXIT_CTD_target'
baseline estimation	breslow
number of observations	252086
number of events observed	88764
partial log-likelihood	-1030013.90
time fit was run	2021-10-29 21:59:36 UTC

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
CREDITO_BANC	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-1.00	0.32	1.65
CAL_RIESGO_FIN	0.01	1.01	0.00	0.01	0.01	1.01	1.01	77.51	<0.005	inf
NUM_EMPLEADOS	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-0.03	0.98	0.03
TRX_PROM_MES	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-0.57	0.57	0.82
VENTAS_TOTALES_YEAR	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.49	0.62	0.69
MES_PROM_VENTAS	-0.00	1.00	0.00	-0.00	0.00	1.00	1.00	-0.54	0.59	0.76
BAFUNDUS	0.00	1.00	0.00	0.00	0.00	1.00	1.00	26.45	<0.005	509.82
sunewtec	0.79	2.21	0.01	0.78	0.81	2.18	2.25	99.05	<0.005	inf
sunowjob	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.13	0.89	0.16
suowners	0.00	1.00	0.00	0.00	0.00	1.00	1.00	2.33	0.02	5.66
suyr5job	0.00	1.00	0.00	0.00	0.00	1.00	1.00	7.32	<0.005	41.90

Concordance	0.62
Partial AIC	2060049.79
log-likelihood ratio test	15460.05 on 11 df
-log2(p) of ll-ratio test	inf

De la tabla 9, observamos que este modelo tiene una confiabilidad de 0,62 pero además de esta métrica identificamos que hay algunas variables que son significativas de acuerdo al valor p (el valor p ayuda a diferenciar resultados que son producto del azar del muestreo,

de resultados que son estadísticamente significativos) (Wasserstein R, 2016) para determinar la supervivencia de la empresa, por ejemplo, la variable CAL_RIESGO_FIN que determina la calificación del riesgo financiero de la pyme está relacionada con la probabilidad de supervivencia de la empresa, como también lo son BAFUNDS que determina los fondos que tiene la empresa como patrimonio, sunewtec (inversión en nuevas tecnologías para la empresa) y suyr5job (cantidad de empleos a generar en 5 años) y esta significancia es con una confiabilidad del 95%.

Análisis de los modelos de clasificación con las variables numéricas.

A continuación, se observa los modelos de clasificación que se entrenaron con el conjunto de variables tipo numéricas.

Tabla 10: Modelos de clasificación entrenados con las variables numéricas.

Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
CatBoost Classifier	0.7238	0.7382	0.4184	0.6775	0.5173	0.3387	0.3580	44.5440
Extreme Gradient Boosting	0.7179	0.7341	0.4464	0.6466	0.5281	0.3364	0.3482	12.3990
Light Gradient Boosting Machine	0.7104	0.7302	0.4856	0.6148	0.5426	0.3350	0.3400	1.5840
Random Forest Classifier	0.7016	0.7140	0.4981	0.5932	0.5415	0.3229	0.3256	26.5270
Extra Trees Classifier	0.6924	0.7039	0.5055	0.5740	0.5376	0.3085	0.3099	38.5060
Gradient Boosting Classifier	0.6877	0.7136	0.5906	0.5551	0.5723	0.3268	0.3272	20.8360
Ada Boost Classifier	0.6869	0.7053	0.5384	0.5598	0.5488	0.3093	0.3095	6.0850
Quadratic Discriminant Analysis	0.6850	0.6650	0.2529	0.6381	0.3621	0.2019	0.2401	0.7660
Ridge Classifier	0.6647	0.0000	0.6059	0.5225	0.5611	0.2922	0.2944	0.4470
Linear Discriminant Analysis	0.6647	0.6708	0.6059	0.5225	0.5611	0.2923	0.2944	0.5990
Naive Bayes	0.6511	0.5304	0.0240	0.6980	0.0463	0.0234	0.0797	0.4220
Decision Tree Classifier	0.6240	0.5972	0.5055	0.4706	0.4874	0.1911	0.1914	1.3980
Logistic Regression	0.5713	0.5158	0.3120	0.4720	0.3049	0.0317	0.0564	1.4090
K Neighbors Classifier	0.5148	0.5016	0.4550	0.3550	0.3988	0.0024	0.0025	1.2460
SVM - Linear Kernel	0.4843	0.0000	0.5586	0.4914	0.3390	0.0023	0.0124	9.4810

De la tabla 10, observamos que el mejor modelo en todas las métricas es el **Catboost Classifier.** Si vemos el accuracy inicialmente está con 0,72.

Ajuste de hiper parámetros para el Catboost Classifier

Con el objetivo de mejorar el modelo seleccionado, ajustamos los hiper parámetros y en la siguiente tabla observamos los resultados. Nuevamente observamos que tenemos el mismo valor, solo ha variado los decimales sin embargo el modelo se sostiene en 72% en el accuracy.

Tabla 11: Ajuste del modelo para mejorar la confiabilidad de las predicciones

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7211	0.7340	0.4172	0.6698	0.5142	0.3331	0.3514
1	0.7225	0.7380	0.4214	0.6718	0.5179	0.3372	0.3553
2	0.7211	0.7373	0.4224	0.6670	0.5172	0.3348	0.3521
3	0.7206	0.7419	0.4180	0.6678	0.5141	0.3323	0.3503
4	0.7200	0.7319	0.4117	0.6694	0.5099	0.3291	0.3482
5	0.7230	0.7324	0.4221	0.6729	0.5188	0.3383	0.3565
6	0.7264	0.7401	0.4255	0.6814	0.5239	0.3460	0.3649
7	0.7207	0.7370	0.4157	0.6695	0.5129	0.3319	0.3504
8	0.7225	0.7399	0.4220	0.6716	0.5183	0.3374	0.3554
9	0.7274	0.7429	0.4298	0.6818	0.5273	0.3493	0.3676
Mean	0.7225	0.7375	0.4206	0.6723	0.5174	0.3369	0.3552
SD	0.0024	0.0036	0.0049	0.0050	0.0049	0.0060	0.0061

Creación del modelo Catboost Classifier para las variables numéricas

Para finalizar, creamos el modelo con los hiper parámetros ajustados y realizamos 10 iteraciones.

Tabla 12: Creación del modelo final que predice la supervivencia de la EMPRESA

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7219	0.7351	0.4145	0.6735	0.5132	0.3336	0.3529
1	0.7267	0.7395	0.4204	0.6853	0.5212	0.3450	0.3651
2	0.7215	0.7365	0.4209	0.6690	0.5167	0.3351	0.3528
3	0.7237	0.7395	0.4183	0.6773	0.5172	0.3385	0.3578
4	0.7209	0.7335	0.4092	0.6737	0.5091	0.3300	0.3501
5	0.7250	0.7342	0.4169	0.6820	0.5175	0.3404	0.3606
6	0.7261	0.7417	0.4249	0.6809	0.5232	0.3452	0.3641
7	0.7210	0.7380	0.4161	0.6702	0.5134	0.3325	0.3511
8	0.7241	0.7404	0.4159	0.6799	0.5161	0.3385	0.3585
9	0.7273	0.7436	0.4268	0.6834	0.5254	0.3482	0.3672
Mean	0.7238	0.7382	0.4184	0.6775	0.5173	0.3387	0.3580
SD	0.0023	0.0032	0.0049	0.0054	0.0047	0.0057	0.0059

Debate

Se realizaron ajustes en los hiper parámetros y tras varias iteraciones de entrenamiento en el modelamiento la precisión se sostuvo en un nivel de 72%, lo que significa que se tiene un margen de error de 28% en las predicciones realizadas con los datos que se tienen, por lo tanto, los resultados para cada conjunto de variables exigen una exploración de otros factores, que pueden ser activos intangibles que afectan la predicción de éxito de la empresa.

La aplicación de modelos de clasificación para diferentes tipos de conjuntos de variables fue importante dado que, al analizarlos por separado se pudo observar el efecto real que tiene cada una de ellas sobre la variable objetivo.

Cuando se realizó la comparación del modelo de regresión Cox para ambos conjuntos de variables, identificamos que los resultados son similares cuando se identificaron las variables significativas sin embargo el poder predictivo de este modelo no fue suficiente y mejor para ganarle a los modelos de clasificación. Este modelo de regresión Cox nos sirve para determinar cómo puede cambiar la probabilidad de supervivencia a través del tiempo.

Los procesos analíticos que se realizaron en el proyecto están enfocados en el conocimiento de las relaciones existentes de los diferentes tipos de variables (binarias o numéricas) con la variable objetivo de tipo binaria: si la empresa sobrevive o no en el mercado. En este caso no se realizaron relaciones no lineales, sino con la relación logarítmica que está intrínsecamente en el modelamiento estableciendo relaciones directas con la variable objetivo.

La situación más importante en el estudio realizado es la identificación de la influencia que tiene cada conjunto de variables en la predicción de la supervivencia y además, determinar con el experto la coherencia de los resultados.

Por ello, al tener estos resultados nos damos cuenta de que las metodologías de *Machine Learning* no pueden dar todas las respuestas, siempre es muy importante tener la asesoría del experto en el tema, en este caso en emprendimiento. Así pues, estos análisis son un complemento potencial para este campo de estudio, pero no reemplazaría la investigación convencional.

Conclusiones

La generación de emprendimientos ha estado en constante crecimiento en los últimos años, diferentes razones políticas, sociales y económicas llevan a las personas a generar sus propios recursos para impactar la sociedad o de alguna manera enfrentar el desempleo, sin embargo, la proporción de emprendedores exitosos es muy baja y sigue decreciendo.

Observamos que más del 90% de las empresas no alcanzan a llegar a los 5 años de vida en el mercado y que muchas de ellas no entienden el por qué fracasaron.

Por ello, identificamos que la supervivencia de las empresas está relacionada a dos diferentes tipos de variables: de negocio (tangibles) y de tipo psicológico (intangibles).

Adicionalmente, el factor del apoyo del Estado - País es importante puesto que si se proporcionan ayudas a estas personas serían piezas fundamentales en la economía de un país, pero si no se generan estas ayudas económicas o se realizan estrategias publicitarias sobre cifras positivas de emprendimiento se multiplicarán los fracasos en las empresas. Así mismo, iniciar un proyecto de emprendimiento con un buen equipo de trabajo puede dar la madurez necesaria para sobrevivir en el mercado.

Las variables de negocio son de importancia en un emprendimiento, pero algunas de estas se estabilizan mucho después de iniciar un proyecto emprendedor; como lo son, ventas mensuales totales, ventas totales anuales, transacciones promedio mes. Pero, observamos que tener una visión clara sobre cuántos empleados se tendrán dentro de los próximos 5 años es un factor clave de supervivencia y además tener una claridad en la toma de decisiones sobre cuánto es la capacidad de endeudamiento en un banco y calificación del riesgo financiero es óptima también son factores claves para que el negocio crezca y permanezca prestando servicios para la comunidad.

El uso de *Machine Learning* orientado a aprendizaje supervisado con modelos de clasificación nos da una visión amplia sobre el poder predictivo que tienen diferentes tipos de variables sobre la supervivencia de las empresas en el mercado.

Finalmente, se observa la importancia de construir modelos considerando variables del negocio tanto tangibles como las variables de negocio intangibles, para relacionar el contexto financiero con el contexto del negocio. Por ejemplo, una empresa que valora y desea centrar su negocio a la fidelización de los clientes, convirtiendo su empresa en modelos aspiracionales, es posible que el análisis con datos tangibles refleje un positivo falso si lo comparamos con el análisis con datos intangible.

Por esta razón, nos motiva, para trabajos futuros, diseñar modelos que relacionen las variables de negocio tangibles con las variables de negocio intangibles, que facilite la consideración de percepciones y valores no estructurados con variables tangibles.

Referencias

Aguilar, J., RamÌrez, N., - Hernandez, C. (2011). La entrada al mercado de las microempresas informales en México y la relación con su expectativa de vida. Revista Internacional de administración y finanzas,4(4),1-14.

Sunil Ray, (2017) CatBoost: A machine learning library to handle categorical (CAT) data automatically

Cader, H., & Leatherman, J. (2011). Small Business Survival and Sample Selection Bias. Small Business Economics,37(2),155-165.

Antoni Baum (2021) “bayesian-hyperparameter-optimization-with-tune-sklearn-in-pycaret”

Glennon, D., - Nigro, P. (2005). Measuring the default risk of small busi-ness loans: A Survival analysis approach. journal of money Credit and Banking, (37(5),923-947.

Headd, B. (2003). Redefining Business Success: Distinguishing between Closure and failure. Small business Economics,21(1),51-61.

Lancaster, T. (1972). A Stochastic Model for the Duration of a Strike. journal of the oyal Statistical Society,135(2),257-271.

Manuel A. Pulido Cayuela. (2008). Generación de números aleatorios, Tema 1. Pag1-18

Ortiz, M. (2013). El fracaso de la microempresa relacionado con las caracterÌsticas individuales del propietario: un estudio empÌrico en Rep ́blica Dominicana.Faedempresa International review,2(3),39-48.

Parra, J. (2011). Determinantes de la probabilidad de cierre de nuevas empresas en Bogotá, Facultad de Ciencias Económicas Investigación y Reflexión,19(1). 27-53.

Santana, L. (2014). Costo de capital de micro y pequeñas empresas en Bogotá· D.C., una aproximación con modelos de probabilidad. Congreso internacional de contabilidad y finanzas, desafíos en los mercados emergentes. 104-14.

Wagner, J. (2013) Exports, imports and firm survival: first evidence for manufacturing enterprises in Germany. review of World Economics / Weltwirtschaftliches Archiv, 149(1), 113-130.

Lin, S., Ansell, J., - Andreeva, G. (2012). Predicting default of a small bu-siness using different definitions of financial distress. The Journal of the operational research Society,63(4),539-548.

Kiefer, N. (1998). Economic Duration Data and Hazard Functions. journal of Economic Literature,26(2),646-679

Amorós, J. y Guerra, M. (2008). GEM, Global Entrepreneurship,Monitor: Reporte Nacional de Chile 2008. Ediciones,Universidad del Desarrollo. Santiago de Chile.

Amorós, J. y Guerra, M. (2009). GEM, Global Entrepreneurship,Monitor: Reporte Nacional de Chile 2009. Ediciones,Universidad del Desarrollo. Santiago de Chile.

Alind Gupta, 2020, Extra Tree Classifier for feature selection

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. Psychological Methods, 2(2), 131–160

N.V Chawla, K.W Bowyer, L.O Hall & W.p Kegelmeyer, 2002 SMOTE: Synthetic Minority Over-sampling Technique

Fernando López, 2021, SMOTE: Synthethic Data Augmentation for Tabular Data

Jahir A. Gutiérrez O.& Elimeleth Asprilla M.& José M.Gutiérrez L ,2014: Emprendimiento e investigación en la escala de formación profesional y la innovación empresarial en Colombia

Roberto González Ortiz, Alejandra Zúñiga Álvarez,2011: Método CEPCES para la evaluación del potencial emprendedor.

Cámara de comercio de Medellín,2019: “Crece la pyme en la base empresarial en Antioquia”.

Confecámaras,2017: “Determinantes de la supervivencia empresarial en Colombia”.

Confecámaras,2019: “4,2% aumentó la creación de empresas durante el primer semestre del 2019”.

Global Entrepreneurship Monitor - GEM, 2018: “Informe GEM España”.

IARA - Consulting Group,2018: “¿Por qué el 70% de las empresas en Colombia fracasan en los primeros cinco años?

Bosma, N., Hill, S., Ionescu-Somers, A., Kelley, D., Levie, J., y Tarnawa, A. (2020). Global Entrepreneurship Monitor 2019/2020 Global Report. London Business School. London: Global Entrepreneurship Research Association.

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A. F., Wilson, J. F., Agakov, F., Navarro, P., &Haley, C. S. (2015). Aplicación de la selección de características de alta dimensión: evaluación para predicción genómica en el hombre. Informes científicos, 5, 10312. https://doi.org/10.1038/srep10312

Bosma, N., Hill, S., Ionescu-Somers, A., Kelley, D., Levie, J., &Tarnawa, A. (2020). Global Entrepreneurship Monitor 2019/2020 Informe Global. Escuela de Negocios de Londres. Londres: Asociación Global de Investigación del Emprendimiento.

Breiman,L. (2001). Bosques aleatorios. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brixy, U., Sternberg, R., & Stüber, H. (2012). La selectividad del proceso emprendedor. Diario de Gestión de Pequeñas Empresas, 50(1), 105-131. https://doi.org/10.1111/j.1540-627X.2011.00346.x

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C.R., &Wirth, R. (2000). CRISP-DM 1.0: Guía de minería de datos paso a paso.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Los elementos del aprendizaje estadístico: minería de datos, inferencia y predicción, segunda edición (2ª 2009, Corr. 9ª impresión 2017 ed.). El springer.

Hundt, C., & Sternberg, R. (2016). Explicación de la creación de nuevas empresas en Europa desde una perspectiva espacial y de tiempo: Un análisis multinivel basado en datos de individuos, regiones y países. Artículos en Ciencias Regionales, 95(2), 223-257. https://doi.org/10.1111/pirs.12133

Reynolds, P., Bosma, N., Autio, E., Hunt, S., De Bono, N., Servais, I., Lopez-Garcia, P., &Chin, N. (2005). Monitor global de emprendimiento: Diseño e implementación de la recopilación de datos 1998-2003. Economía para pequeñas empresas, 24(3), 205-231. http://doi.org/10.1007/s11187-005-1980-1

Sokolova M., Japkowicz N., & Szpakowicz S. (2006). Más allá de la precisión, F-Score y ROC: Una familia de medidas discriminatorias para la evaluación del rendimiento. En: A. Sattar, & B. Kang (Eds.) IA 2006: Avances en Inteligencia Artificial. Berlín: Springer. https://doi.org/10.1007/11941439_114

Patrik Shukla (2020) Guide survival Analysis Python

Anarag Pandey (2019) Análisis de supervivência:Implementación em Python.

Wasserstein RL, Lazar NA (2016). «The ASA's statement on p-values: context, process, and purpose». The American Statistician. doi:10.1080/00031305.2016.1154108.