PROYECTO PILOTO DE ANALÍTICA DE DATOS - SUPERVIVENCIA EMPRESARIAL

Marzo 2020

7

DATA-POP ALLIANCE

mit media lab

DNP Departamento Nacional de Planeación

iNNpulsa Colombia

Producido por un equipo compuesto por:

Maria Antonia Bravo, Andrés Lozano

Basado en documentación de EyS Soluciones Empresariales IT SAS para iNNpulsa Colombia

Coordinado por Emmanuel Letouzé

Bajo la supervisión general de Alex Pentland

Diagramación editorial por Paola Caile

Marzo 2020

Marzo 2020
Versión revisada y ajustada

ÍNDICE

Glosario

Arquitectura de datos: arquitectura que describe cómo se recolecta, almacena, transforma, distribuye y consume los datos dentro de una organización o empresa. Incluye reglas que conciernen la estructuración de formatos, incluyendo bases de datos y sistemas de archivos, además de incluir reglas sobre cómo se vinculan datos a través de sistemas para el uso de estos a través del proceso comercial (DalleMule y Davenport 2017). Entre los elementos de una arquitectura de datos se encuentra el acceso a los datos, almacenamiento en la nube, procesamiento de datos, redes, encriptación, hosting entre otros (Knight 2018).

Servicios de AWS: Amazon Web Services (AWS) es una plataforma de servicios basados en la nube que incluye recursos para el cómputo, almacenamiento, y análisis de datos, incluyendo herramientas para desarrolladores y aplicaciones empresariales (“AWS” s/f).

Data lake: “un data lake es un repositorio de almacenamiento que contienen una gran cantidad de datos en bruto y que se mantienen allí hasta que sea necesario. A diferencia de un data warehouse jerárquico que almacena datos en ficheros o carpetas, un data lake utiliza una arquitectura plana para almacenar los datos” (“Data Lake: Definición, Conceptos Clave y Mejores Prácticas” s/f).

Metadatos: conjunto de datos que describe y proporciona información sobre otros datos. En otras palabras, datos acerca de los datos (“Metadatos, Definición y Características” s/f).

Introducción

El objeto de este documento es describir el proceso de realización de un piloto de analítica de datos, concretado en el marco del contrato suscrito entre iNNpulsa Colombia y el Massachusetts Institute of Technology (MIT). Para este piloto, se realizó la contratación de la empresa EyS Soluciones Empresariales, cuyo objetivo fue crear y ejecutar un modelo de analítica de datos predictivo y diagnóstico de los factores determinantes de la supervivencia y exportación de las empresas colombianas. .

Para este análisis, el piloto hizo uso principalmente de Survival Analytics, familia de métodos estadísticos y de aprendizaje de máquinas para estimar la duración y analizar el tiempo de eventos. La mayor ventaja de utilizar métodos de Survival Analytics es la posibilidad de estimar la probabilidad de que suceda el evento y también el momento de ocurrencia. Específicamente, se utilizó Survival Analytics para 1) estimar la probabilidad de supervivencia de una empresa antes de cierto momento, 2) estimar la tasa de riesgo de que el evento de desaparición ocurra súbitamente, 3) estimar la probabilidad de que una empresa inicie actividades exportadoras y por último 4) la tasa de riesgo de que el evento de exportación ocurra súbitamente. Para complementar estos métodos, se utilizaron también otros modelos estadísticos paramétricos y no-paramétricos, y de aprendizaje de máquinas.

Entender con mayor precisión la incidencia que factores sociales, económicos, y empresariales tienen sobre la supervivencia y la vocación exportadora de empresas colombianas, es clave para generar políticas públicas informadas que reflejen la realidad contextual y cambiante del país. Este proyecto además de ser un insumo para políticas públicas a futuro, es incluso, un ejemplo claro del alcance y tipo de resultados que es posible obtener al adelantar proyectos de analítica en el marco de una estrategia de Big Data para el estado. Además, es prueba del compromiso del gobierno en afianzar alianzas con el sector privado y con talento local para el desarrollo de tecnologías y afianzamiento de los cambios precipitados por la Cuarta Revolución Industrial.

Este proyecto representa el avance del estado colombiano hacia una administración pública que utiliza los datos para crear mejores políticas, para retroalimentar procesos existentes, y para crear nuevos procesos que pueden guiar a los tomadores de decisiones hacia decisiones mejor informadas. De igual manera, el uso de Survival Analytics para estudiar la supervivencia empresarial presenta un trabajo de vanguardia, pues hasta la fecha, no se han realizado trabajos académicos aplicando Random Forest y Random Survival Forest a problemáticas de esta naturaleza.

Objetivos específicos del piloto

Específicamente, los objetivos plasmados para este proyecto, incluyeron:

1. La probabilidad de que una empresa sobreviva de acuerdo con las variables que sean identificadas por el Consultor que caracterizan el contexto histórico y situación actual de las empresas.

2. La probabilidad de que una empresa comience a exportar de acuerdo con las variables que sean identificadas por el Consultor que caracterizan el contexto histórico y la situación actual de la empresa.

Este documento recuenta el proceso de elaboración del proyecto incluyendo el inventario de datos utilizados, la descripción técnica del análisis, resultados cualitativos y cuantitativos, así como retos y recomendaciones para entidades públicas interesadas en adelantar proyectos similares. El contenido de este es un resumen de la documentación elaborada por los consultores contratados para llevar a cabo este proyecto. Para conocer más sobre este proyecto, sus autores, y sus resultados en detalle, ver aquí.

I. Inventario de datos utilizados

1) Registro Único Empresarial y Social (RUES)*, base de datos administrada por Confecámaras que incluye los registros de todos las Cámaras de Comercio en Colombia (2000-presente). Contiene información detallada de las empresas registradas en las Cámaras de Comercio del país. Una de las limitaciones de esta fuente de datos es la inclusión de empresas registradas y declaradas, mas no incluye empresas de papel o fachada. Igualmente, no se toman en consideración fusiones o adquisiciones entre compañías, lo que hace de esta fuente una base de datos no exhaustiva de las empresas colombianas. Los indicadores financieros incluidos en esta base de datos son declarativos, incluyendo ingresos y capital. Más de 12 millones de registros tipo panel en 90 variables.

2) Importaciones DANE-DIAN*, base de datos que reúne información del comportamiento de las importaciones. Contiene información desde 2008 al 2019. Solo incluye información de bienes importados, no de servicios.

3) Exportaciones DANE-DIAN*, base de datos que reúne información del comportamiento de las exportaciones. Contiene información desde 2008 al 2019. No incluye información de exportación de servicios.

4) Sistema de Matrículas Estudiantil (SIMAT) es una de las bases de datos de educación del Ministerio de Educación, incluye información sobre las matrículas de estudiantes de instituciones oficiales,. Más de un millón de registros tipo panel en 11 variables, no obstante menos de mil registros coinciden con representantes legales de las empresas, ya que no se consideran instituciones privadas o de educación superior.

Para ver las variables utilizadas en el marco de este proyecto, ver Anexo A.

II. Descripción técnica del análisis efectuado

Esta sección describe el análisis efectuado desde la recepción de los datos, la arquitectura creada para llevar a cabo el mismo, hasta el tratamiento y el procesamiento de los datos para el desarrollo del proyecto piloto de analítica de datos.

A. Arquitectura

La arquitectura de datos es uno de los pilares principales para desarrollar proyectos de analítica, pues define no sólo la forma en la que se almacenan los datos, pero también incide en la configuración de las bases de datos, el poder de procesamiento de estos datos, así como las necesidades de mantenimiento y calidad de estos. La arquitectura utilizada y desplegada por los consultores se centró en Amazon Web Services (S3 Standard). Esta arquitectura cumple con las garantías de seguridad necesarias (certificadas internacionalmente) y está alineada con estándares y mejores prácticas vigentes para soluciones de Big Data, Analítica Avanzada e Inteligencia Artificial.

Cuadro 1. Arquitectura de datos: condiciones para obtener alto valor operativo para modelos con enfoque institucional o corporativo

- Serverless: La plataforma de procesamiento, almacenamiento e interconexión de red debe prescindir de servidores persistentes o permanentes; migrando las cargas de trabajo a infraestructuras efímeras. De esta forma, los ciclos de vida están supeditados a los periodos de procesamiento requeridos, según la dinámica de análisis requerida. Igualmente, la asignación de recursos de cómputo se ajusta automáticamente para procesar las respuestas a las preguntas de negocio dentro de los márgenes de tiempo requeridos por el negocio.

- Escalabilidad sostenible: La arquitectura debe ser capaz de escalar ágilmente ante la demanda de volumetrías de datos adicionales o tiempos de ejecución más veloces. Tal que los modelos desplegados no necesiten ser reconfigurados para soportar crecimientos en los recursos de procesamiento, almacenamiento e interconexión de los servidores o ampliaciones del número de nodos (servidores de procesamiento).

- Autonomía: Los modelos deben estar en la capacidad de ser activados, re-entrenados, re-evaluados y desplegados productivamente sin ningún tipo de intervención manual. La orquestación de servicios en la nube y ejecución de modelos es delegada en la programación de tareas conociendo los tiempos y frecuencias requeridas por el negocio.

- Uso eficiente: Los recursos de cómputo, almacenamiento e interconexión empleados para el despliegue de los modelos se realiza bajo un esquema por demanda, es decir, los servicios de nube que hacen parte de la arquitectura son activados mediante eventos previamente definidos, tales como: actualización de las fuentes de datos, actualización de los modelos, perfeccionamiento de los modelos, entre otros. Las eficiencias generadas no solamente son de índole tecnológico, sino también financieras.

Tomado de: “Entregable 3A - Metodología de Analítica Avanzada de Datos” (Chaparro et. al, 2020)
Fuente original (Yao 2017, Provost 2013).

La arquitectura utilizada incluye los siguientes servicios:

Servicio en AWS: S3 Standard
Gestión de Acceso e Identidad: usuarios, grupos y roles
Ambientes soportados: i) ‘as-is’ para datos en su estado original. ii) ‘in-proc’ para datos en proceso de transformación o mejoramiento. iii) ‘staging’ para estados preliminares de integración o consumo de datos por otro servicio y iv) ‘output’ estado final de preparación, transformación o inferencia.

Este servicio fue utilizado en la creación de un data lake, o repositorio de datos en bruto.

Servicio en AWS: Sagemaker
Nombre instancias: ml.c5.4xlarge
Características instancias: 16 vCPU, 32GB RAM, 10 Gbps tasa transferencia de red
Sistema Operativo: Amazon Linux 2018.03
Software especializado: Python 3.6

Este servicio de programación de tareas fue utilizado en los procesos de preparación, modelación y evaluación de los datos, pues permite de manera automática activar instancias virtuales con preconfiguraciones determinadas para el análisis de los datos. Esto permite analizar los datos de manera más eficiente y procesar datos en “autopiloto”. De igual manera, contribuye en el proceso de transformar datos brutos en información; esta información es después almacenada en el data lake.

Servicio en AWS: Glue Data Catalog
Este servicio permite la creación de un catálogo de fuentes para la gestión de los datos y los metadatos de los cubos de información generados por los modelos de preparación de los datos, así como de los modelos de análisis diagnósticos y predictivos. El catálogo de fuentes incluye el nombre, tipo de fuente, tamaño y formato de la metadata.

Servicio en AWS: Athena
Este servicio permite la consulta (*query*) de la información almacenada en formatos de tablas estructuradas en cubos de información a través de SQL (lenguaje de consulta y acceso a bases de datos relacionales).

B. Procesamiento y reingeniería de datos

1. Extracción de fuentes de datos

La entrega de los datos a los consultores se realiza de manera anonimizada mediante SharePoint, donde se transfiere 2.2 GB conteniendo 197 columnas y 2.6 mil millones de filas de las bases de datos descritas arriba. Tras la recepción de estos y la creación de un data lake en la nube, se caracterizan a primer nivel las fuentes de datos suministradas través de la definición de los metadatos, y se obtiene un listado completo de las variables que componen cada fuente de datos con su correspondiente definición, tipo, formato, regla de validación, valores de dominio, valor por defecto, reglas de negocio, entre otras. Adicionalmente, se analiza la volumetría de los datos recibidos, incluyendo el tamaño de los datos en bruto, la cantidad de registros y diferenciales de tamaño por hora, día, semana o mes.

Previo a la integración de las fuentes de datos, se anonimizan las variables con información de carácter personal[1], utilizando técnicas de hashing SHA1.

2. Diagnóstico de calidad de los datos

En esta fase se verifica la integridad y consistencia de las fuentes y sistemas de información a través de tres métricas de calidad de los datos: precisión, completitud y duplicación. Precisión incluye medir que los diferentes valores existentes para cada variable sean consistentes a través de la base de datos en su función y definición. Completitud incluye examinar la cantidad de registros nulos, vacíos o inválidos. Duplicación incluye remover los registros u observaciones de cada fuente cuyos valores por cada variable sean idénticos para asegurarse una entrada por observación.

A través de este diagnóstico se hace posible priorizar las variables o fuentes con necesidad de reingeniería o enriquecimiento de metadatos y datos.

3. Reingeniería de datos

La reingeniería de datos tiene como objetivo estandarizar las definiciones y formatos de los datos, depurar valores o información inválida, homologar variables encontradas en diferentes fuentes de información y la limpieza en general de los datos, entre otras. Este proceso incluye

[1] Las variables anonimizadas incluyen el NIT, Razón Social, Representante Legal y las variables borradas incluyen Dígito Verificación (RUES, DIAN), Dirección Fiscal (RUES, DIAN), Teléfono Fiscal (RUES, DIAN), Nombre Revisor Fiscal (RUES, DIAN), Dirección Comercial (RUES), Teléfono Comercial (RUES), Correo Electrónico (RUES), NIT Declarante DIAN, Mineducación), Razón Social SHA (Mineducación), Representante Legal SHA (Mineducación) y Número Identificación Representante Legal SHA (Mineducación).ron

por ejemplo, la eliminación de los caracteres inválidos existentes y transformación de celdas como R$%gim/n a “régimen”, o convertir todos los datos a minúscula.

4. Integración y agregación

Tras la reingeniería de los datos para cada una de las bases, se ensambla un cubo de información incluyendo estas fuentes en formato tipo panel[2], el cual tiene una estructura principal de orden temporal. Esto se hace al lograr la unión entre las diferentes fuentes, utilizando como llave principal el NIT de las empresas para esta intersección y complemento. Lo anterior, permite reunir en una sola base de datos todas las variables a estudiar para cada una de las empresas, resumiendo la información en valores únicos con múltiples entradas transaccionales o temporales, incluyendo múltiples registros por empresa. Adicionalmente, en este cubo de datos, se establecen con claridad las variables con el rol diagnóstico o predictivo y las variables objetivo, tales como la duración y evento de desaparición.

El primer nivel de análisis descriptivo se realiza sobre esta base de datos tipo panel, al producir métricas estadísticas tales como mínimos, máximos, medias, medianas, entre otras. Para las variables categóricas se realizan distribuciones de frecuencia para “conocer la densidad de sus valores de dominio”.

Este cubo de información contiene información de 5.4 millones de empresas, con 120 atributos por empresa. El Anexo A incluye información sobre las operaciones de agregación aplicadas a estas variables.

5. Reducción de dimensionalidad

Esta etapa comprende el proceso de identificar, corregir o descartar las covariantes o variables que comprometen “la convergencia o estimación exitosa de los modelos”, pues consiste en remover los covariantes que contienen información redundante entre sí. La reducción de dimensionalidad busca eliminar el efecto de la codependencia linear, para así promover la confiabilidad en los determinantes del modelo y sus pesos respectivos, particularmente en el proceso de identificar factores determinantes de supervivencia empresarial y vocación exportadora. Para ello, se automatizó el cálculo de los coeficientes de correlación de Spearman y se descartaron aquellas variables por encima de ± 0.8. En este proceso se descartaron 54 variables[3].

[2] Una base de datos tipo panel es un tipo de base de datos que contiene datos longitudinales que involucran la medición de una o varias variables a través del tiempo. Estas contienen observaciones de diferentes fenómenos a través del tiempo para el sujeto de observación (individuos, firma, etc).
[3] Entre estas se encuentran ‘Municipio fiscal’, ‘Cód. Categoría’, ‘Cód. Lugar Salida’, ‘Dpto. Procedencia’ entre otras.

C. Modelamiento y análisis de datos

1. Métodos no paramétricos

Un modelo estadístico no-paramétrico es aquel cuyas pruebas y modelos no dependen de una distribución subyacente en los datos, en donde la inferencia de los datos no se hace con una distribución probabilística conocida (eg. distribución normal), sino que se trabaja con la distribución de los datos observados. Esto quiere decir que pueden ser aplicados incluso si las condiciones para la validez paramétrica no se han cumplido (i.e, distribución). Los modelos no-paramétricos de análisis de supervivencia utilizados en el marco de este proyecto piloto incluyen el Método Kaplan-Meier y Método Nelson-Aalen.

Cuadro 2. Ventajas y Desventajas - Métodos No Paramétricos

Ventajas	Desventajas
● Alta eficiencia cuando se desconocen o son difíciles de estimar las funciones de distribución de supervivencia reales ● Tiempos de procesamiento y estimación son bajos, con poca demanda de recursos de cómputo	● Alta dificultad en la interpretación de resultados, lo cual puede llevar a estimaciones inexactas ● Solo considera duración y el evento de desaparición como variables de input en la estimación de las probabilidades de supervivencias y tasas de riesgo

Tomado de: “Entregable 3A - Metodología de Analítica Avanzada de Datos” (Chaparro et. al, 2020)

Método Kaplan-Meier[4]

Este modelo permite conocer el tiempo de vida (duración) y evento de vida o muerte por cada empresa. Para este modelo se requiere saber únicamente el tiempo de vida (duración) y evento de muerte (si existe) por cada empresa. La estimación de las probabilidades de supervivencia o de primera exportación es la razón entre la cantidad de empresas desaparecidas (o exportadoras) y la cantidad de empresas existentes, y por tanto en riesgo de desaparición (o en ‘riesgo’ de primera exportación) en el tiempo de interés, resultando en una probabilidad de 0 a 1 de supervivencia o de exportación por diferentes momentos de interés bajo estudio. Por ejemplo, con este método, se estima que la probabilidad de que una empresa empiece a exportar desde Medellín a cinco años es de 89.5%.

Método Nelson-Aalen

Este modelo permite estimar la función de riesgo de desaparición de una empresa, dado a través de los cambios súbitos en la probabilidad de supervivencia empresarial. Este método estima la función de riesgo acumulada, en otras palabras, la tasa de muerte instantánea antes del tiempo de interés. Este modelo estima unas tasas de riesgo (siempre valores positivos) por

[4] Para más información sobre el método Kaplan Meier ver Anexo C - Fórmulas.

cada momento de interés bajo estudio (pueden ser meses o años) para las empresas. Por ejemplo, para el caso colombiano, la tasa de riesgo en el primer año es más acelerada, dado que la diferencia entre las probabilidades de desaparición entre el primer y segundo año son de ocho puntos porcentuales. A través del tiempo esta tasa de riesgo se desacelera.

2. Métodos paramétricos

Los métodos paramétricos son aquellos que asumen y dependen de una distribución estadística conocida. Tal como afirman los autores de este proyecto, “el principio de estos modelos es encontrar una función de distribución de probabilidad parametrizable que describa lo más cercano posible la distribución real de supervivencia de las empresas colombianas.” A través de estos métodos, se estiman y contrastan probabilidades de supervivencia y probabilidades de exportación, y tasas de riesgo sobre la totalidad de empresas estudiadas.

Cuadro 3. Ventajas y Desventajas - Métodos Paramétricos

Ventajas	Desventajas
● Fácil interpretación ● Estimaciones más eficientes y precisas cuando los tiempos de supervivencia siguen alguna función de distribución conocida.	● Resultados inconsistentes e inestables cuando las probabilidades de supervivencia reales no se ajustan adecuadamente a las funciones de probabilidad de referencia. ● Solo considera como input al modelo la duración y el evento de desaparición en la estimación de las probabilidades de supervivencias y tasas de riesgo

Tomado de: “Entregable 3A - Metodología de Analítica Avanzada de Datos” (Chaparro et. al, 2020)

Distribución Paramétrica Weibull

La distribución paramétrica Weibull es una distribución de probabilidad continua (en este caso de supervivencia) que determina la probabilidad de que un evento ocurra en el tiempo.[5]

[5] La distribución paramétrica de Weibull está definida por la fórmula . # “Donde , son los parámetros de la distribución paramétrica Weibull, estimados a través del método de ajuste: maximum likelihood estimation (MLE), el cual emplea una muestra del conjunto de datos suministrado para calcular los parámetros a través del método de optimización iterativa de Newton. El tiempo de convergencia y el número de iteraciones requeridas están en función del tamaño de la muestra de datos hasta satisfacer el criterio de parada (stopping criterion) Lipschitz-Hessian xk+1-x*0.5xk-x*2 es decir, el error de la nueva iteración debe ser menor o igual al error cuadrado de la iteración anterior (Cameron 2019). El estudio en cuestión llevó a cabo una estimación exhaustiva, tomando la totalidad de las más de 5 millones de observaciones, en lugar de muestras parciales. Estimados los parámetros , , también se puede generar la función de riesgo acumulada, definida así: H(t)=t donde , son los parámetros de la distribución paramétrica Weibull. Las distribuciones paramétricas o de referencia pueden tomar diferentes formas. En el presente estudio se consideraron: Weibull, Exponencial, Log-normal, Log-Logistic, Piecewise Exponential y Gamma Generalizada”

3. Métodos de Regresión

Los métodos de regresión permiten identificar y priorizar las variables determinantes en la supervivencia y riesgo estimados, tomando en cuenta no solo las variables de duración y evento de supervivencia o desaparición, sino también las demás variables o covariantes presentes en el cubo de información.

Cuadro 4. Ventajas y Desventajas - Métodos de Regresión

Ventajas	Desventajas
● Incorpora en el cálculo de las funciones de supervivencia y riesgo todos los covariantes disponibles, aparte de la duración y el evento de desaparición. ● La estimación o conocimiento de la distribución real de los tiempos de supervivencia no es requerida. ● Posibilita el análisis de covariantes para medir el impacto en las funciones ante la variación controlada de uno o más variables, mientras las demás permanecen estáticas.	● La distribución de las funciones de supervivencia y riesgo resultantes son desconocidas, lo cual puede llevar a dificultades en la interpretación de los resultados. ● Los tiempos de estimación y convergencia son elevados en comparación con los métodos (no-)paramétricos ● La convergencia es altamente sensible a la estructura de la matriz de covariantes, ya que conjuntos de datos desbalanceados entre eventos de muerte y right-*censored* pueden dar lugar a problemas de singularidad matricial y por ende hacer inviable su inversión. ● La demanda de recursos de cómputo es considerable dependiendo de la cantidad de empresas y número de covariantes.

Tomado de: “Entregable 3A - Metodología de Analítica Avanzada de Datos” (Chaparro et. al, 2020)

Método Cox Proportional Hazard

El método Cox Proportional Hazard estima la función de riesgo a partir de los aportes parciales de riesgo de los covariantes, es decir, informando sobre el riesgo que cada variable contribuye al riesgo global del fenómeno de interés. Para más detalle ver las fórmulas de estos modelos en el Anexo C.

Método Tiempo de Falla Acumulado (Accelerated Failure Time)

Este método considera dos distribuciones de tiempos de supervivencia: A y B. Se asume que A corresponde “al conjunto de empresas bajo estudio” y B a “una distribución paramétrica de referencia”. A diferencia del método Cox PH, el efecto de los covariantes en las funciones de

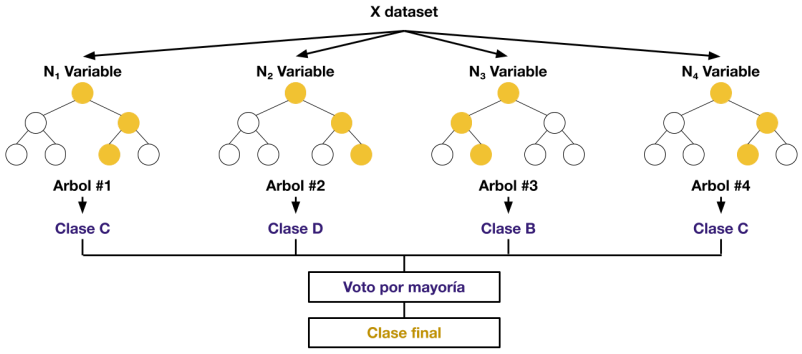
supervivencia y riesgo tienen una incidencia proporcionalmente inversa, en vez una relación lineal.

IV. Random Forest

El Random Forest es un modelo de aprendizaje de máquinas predictivo basado en la construcción de cientos de miles de árboles de decisión[6]. Un árbol de decisión es un algoritmo de aprendizaje supervisado no-paramétrico que construye modelos de clasificación o regresión en la forma de un árbol, donde los datos son divididos en subsets cada vez más pequeños. En palabras simples, el objetivo de un *Random Forest* es utilizar árboles de decisión para crear un modelo predictivo de un valor objetivo (en este caso factores determinantes en la supervivencia y vocación exportadora de las empresas), en donde éste aprende de reglas de decisión simples inferidas de las características de los datos. El modelo aprende de diferentes "tests" realizados sobre los atributos de los datos, y cada uno de los árboles de decisión tienen una capacidad de inferencia limitada.

Sin embargo, y debido al principio fundamental en el cual se basa el random forest - la "ley de los números grandes" - se pondera esta capacidad de inferencia para lograr resultados que de manera conjunta, "superan los niveles de aprendizaje y predicción" de un solo árbol de decisión. Estos árboles de decisión por una parte, son aleatorios, lo "cual ayuda a aumentar la capacidad de aprendizaje del modelo". Por otra parte, por cada nodo de un árbol (o rama), un subconjunto de covariantes es seleccionado aleatoriamente para proseguir con la bifurcación y crecimiento del árbol. La totalidad de los miles de árboles, permite la "generación de diversas reglas de inferencia mientras se mantienen bajo los errores de generalización, es decir, aumenta la precisión de acierto ante la llegada de datos nuevos e inéditos" (Chaparro et. al 2020).

La aplicación de *Random Forests* al análisis de supervivencia empresarial consiste de procesos estándares de modelos de aprendizaje supervisado. Este se realizó a través de cinco validaciones cruzadas (cada una con el 20% total de más de 5 millones de empresas), donde el 80% de los datos se toman para entrenar el modelo, y el 20% restante como prueba.



[6] Los árboles de decisión son un tipo de algoritmo de aprendizaje supervisado utilizados principalmente para problemas de clasificación.

Cuadro 5. Ventajas y Desventajas - Random Forests

Ventajas	Desventajas
● Alto rendimiento de aprendizaje y predicción en la estimación de probabilidades y eventos de supervivencia empresarial, soportado en las capacidades de inferencia propias de técnicas de aprendizaje de máquinas. ● Alta transparencia en la interpretación de la operación interna y resultados del modelo. ● Capacidad de identificar las variables importantes o factores determinantes con un nivel de importancia cuantitativo. ● Comportamiento estable en el uso y demanda de recursos de procesamiento ante grandes cubos de información.	● Designación de importancia prioritaria de variables relacionadas con tiempos y eventos de supervivencia, comprometiendo la identificación de variables con igual o superior impacto en las capacidades de inferencia

Tomado de: “Entregable 3A - Metodología de Analítica Avanzada de Datos” (Chaparro et. al, 2020)

V. Random Survival Forests

Random Forest tiene la limitación de otorgar una alta importancia a variables relacionadas con tiempo y eventos de supervivencia que inciden mayoritariamente en la estimación de probabilidades, restando puntajes de mayor importancia a factores de alto impacto para proyectos de esta índole. Un modelo emergente para contrarrestar estas condiciones restrictivas es *Random Survival Forest (RSF)*, el cual preserva todas las ventajas y beneficios de *random forest*, pero con la capacidad de darle “un tratamiento diferencial a las variables asociadas a tiempo de supervivencia” para lograr maximizar “la pureza en el crecimiento de los árboles” (Chaparro et. al, 2020). Este modelo selecciona dos tercios de las observaciones de las empresas, con sus covariantes, para entrenamiento, y otro tercio para *test*. A diferencia del *random forest*, este modelo maximiza la diferencia en tiempo de supervivencia entre nodos hermanos, brindando la mayor ventaja de utilizar este modelo.

“La priorización de covariantes o identificación de factores determinantes sigue un proceso similar al empleado en random forest. Los valores mayores a 0 son considerados como covariantes o factores con capacidades predictivas, mientras los valores iguales o menores a 0 pueden ser ignorados. Para medir la importancia de un covariante a la vez, las observaciones de prueba son expuestas al conjunto de árboles construidos con las observaciones de entrenamiento. Al encontrar un nodo divisor que coincide con el covariante de turno, se asigna

un nodo hijo aleatoriamente. Dando paso al crecimiento de una colección de árboles con reglas de inferencia aleatoriamente diferentes a aquellas construidas con los datos de entrenamiento. La función de riesgo acumulada de cada árbol es estimada y después promediada con las demás funciones de riesgo de los otros árboles generados. La importancia del covariante es el resultado de la diferencia entre el error de predicción con los árboles ensamblados originalmente y el nuevo ensamble de árboles producido" (Chaparro et. al, 2020).

A diferencia de los métodos paramétricos (Distribución de Weibull), o de métodos que dependen de riesgos proporcionales (i.e, Cox Proportional Hazard), con los *Random Survival Forests* modelar los efectos no-lineales de los covariantes e identificar interacciones entre múltiples variables es más sencillo. Una de las limitaciones de este método es que otorga una alta importancia a variables relacionadas con el tiempo y los eventos de supervivencia, que inciden mayoritariamente en la estimación de eventos de supervivencia, restando puntajes de mayor importancia a covariantes de alto impacto.

Cuadro 6. Ventajas y Desventajas - Random Survival Forests

Ventajas	Desventajas
● Alto rendimiento de aprendizaje y predicción en la estimación de probabilidades y eventos de supervivencia empresarial, soportado en las capacidades de inferencia propias de técnicas de aprendizaje de máquinas. ● Alta transparencia en la interpretación de la operación interna y resultados del modelo. ● Capacidad de identificar las variables importantes o factores determinantes con un nivel de importancia cuantitativo. ● Uso diferenciado de variables relacionadas con tiempos y eventos de supervivencia para aumentar las capacidades predictivas del modelo. ● Convergencia, eficacia y eficiencia limitada por el tamaño o volumetría del cubo de información usando librerías estándar en R y Python. ● Ausencia de soporte para procesamiento multi-nodo y/o multi-núcleo en librerías estándar de R y Python.	● Alto rendimiento de aprendizaje y predicción en la estimación de probabilidades y eventos de supervivencia empresarial, soportado en las capacidades de inferencia propias de técnicas de aprendizaje de máquinas. ● Alta transparencia en la interpretación de la operación interna y resultados del modelo. ● Capacidad de identificar las variables importantes o factores determinantes con un nivel de importancia cuantitativo. ● Uso diferenciado de variables relacionadas con tiempos y eventos de supervivencia para aumentar las capacidades predictivas del modelo. ● Convergencia, eficacia y eficiencia limitada por el tamaño o volumetría del cubo de información usando librerías estándar en R y Python. ● Ausencia de soporte para procesamiento multi-nodo y/o multi-núcleo en librerías estándar de R y Python.

"Entregable 3A - Metodología de Analítica Avanzada de Datos" (Chaparro et. al, 2020)

III. Resultados cualitativos y cuantitativos

Los resultados de este piloto responden a las preguntas acordadas en el marco de este proyecto. Cabe aclarar que los resultados cualitativos y cuantitativos descritos abajo no son exhaustivos de los hallazgos hechos durante este proyecto - más bien, ilustran el tipo de resultados que se obtuvieron por pregunta de acuerdo a los métodos utilizados. Para más detalle sobre los resultados, referirse a los documentos realizados por los consultores.

1) ¿Cuál es la probabilidad de supervivencia empresarial en Colombia?

Según las estimaciones realizadas con el método Kaplan-Meier, la probabilidad de supervivencia empresarial en Colombia, para las empresas estudiadas en el marco de este proyecto, están descritas en la Tabla 1. Estos resultados ponen en evidencia que la probabilidad global de supervivencias disminuye a través de los años, notablemente del primer al décimo año hay una diferencia en la probabilidad de supervivencia de 33.1 puntos porcentuales. De acuerdo a estas probabilidades, al pasar 10 años, 2 de cada 5 empresas colombianas seguirán existiendo.

Tabla 1. Probabilidad de Supervivencia de Empresas Colombianas

Año	Probabilidad de Supervivencia (%)
1	77.6
2	69.5
3	63.4
5	54.3
10	40.5

Sin embargo, estas probabilidades varían dependiendo de las características asociadas a la empresa, incluido el tipo de identificación (como un proxy de persona natural o jurídica), el tipo de organización jurídica, municipio, tamaño empresarial (medido en activos), por vocación importadora y/o exportadora y por actividades económicas (ramas de cuentas nacionales).

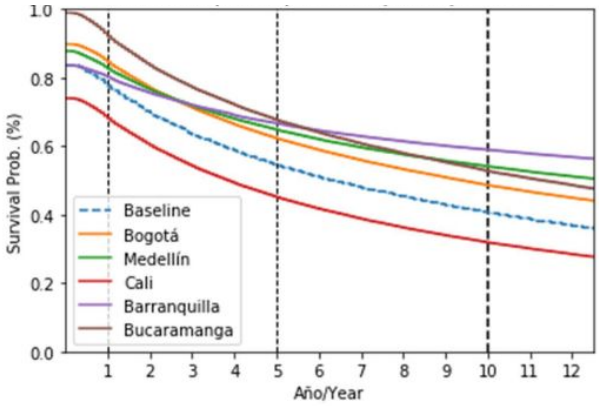
Entre los resultados a destacar se encuentran los siguientes:

- “La supervivencia de las **empresas constituidas como persona jurídica** tienen **mayor probabilidad de supervivencia** con respecto a las de personas naturales en un 6% al primer año y un 7% a los 10 años”
- “Las empresas de **personas naturales y unipersonales** tienen **probabilidades inferiores** en un -3% con respecto a la supervivencia global o línea base, la cual es el valor de la probabilidad de supervivencia considerando el universo completo de empresas. Por otro lado, las **comanditas, sociedades anónimas y limitadas mejoran su supervivencia** en

un +8% para las primeras y hasta más de +25% para las últimas, con respecto a la línea base."

- "**Medellín, Barranquilla y Bucaramanga** evidencian **probabilidades de supervivencia superiores** en el año 1, 5 y 10 hasta 20% más en comparación con la línea base. Mientras, **Bogotá** tiende a **desmejorar su oportunidad de supervivencia** con el paso del tiempo con +9% por encima de la referencia global. En Cali, se observan probabilidades consistentemente **más bajas en -9%**."

Figura 1. Probabilidad de supervivencia por municipio (Kaplan-Meier)



Fuente: Chaparro et. al, 2020

- Para empresas que reportan **activos iguales a 0**, **1 de cada 3 empresas sobrevive** después del décimo año. Las empresas con **activos de más de 50 millones**, "gozan de **probabilidades muy altas de supervivencia**, por encima del 80% incluso para el año 10, transcurrido desde el inicio del periodo del presente estudio."
- Más del **95% de las empresas** estudiadas **no tienen vocación exportadora o importadora** de bienes. Sin embargo, "cabe resaltar que se observa un **aumento de la supervivencia** en 27% para las **importadoras** y +10% en las **exportadoras** para el año 10." Para las **empresas exportadoras e importadoras**, la probabilidad de supervivencia durante los **primeros dos años es 4% por encima de la línea base**; después de este tiempo su comportamiento es igual a la de las empresas sin vocaciones exportadoras o importadoras.
- "Las empresas con **actividades inmobiliarias** exhiben **tasas de supervivencia superiores al 70%** y lejanas a las demás cuentas". "Las empresas asociadas a la **industria manufacturera y actividades profesionales, científicas y técnicas muestran las siguientes mejores tasas** después de las empresas con actividades mobiliarias."

Por otro lado, al utilizar métodos de aprendizaje supervisado, el modelo utilizado logra predecir correctamente el número de empresas que sobrevivirán tanto como las que desaparecerán con un 98.65% de *accuracy.* La Tabla 2 abajo describe la matriz de confusión del modelo, demostrando un alto desempeño en la predicción.

Tabla 2. Matriz de confusión del modelo *Random Forest* - Supervivencia de empresas

	Predicción	
Realidad	Desaparece	Sobrevive
Desaparece	**407796**	7018
Sobervive	8663	**657500**

Es importante mencionar también que en modelo entregado por los consultores a iNNpulsa es posible analizar la supervivencia empresarial según co-variantes (por ejemplo, restaurantes por ciudad) para así observar los cambios en las probabilidades de supervivencia y tasas de riesgo con más de un factor de interés, siendo esto un insumo crucial en la construcción de políticas públicas.

Métodos utilizados: Kaplan-Meier, Nelson-Aalen, Random Forest

2) ¿Cuáles son factores determinantes de la supervivencia empresarial en Colombia?

Los factores determinantes de la supervivencia empresarial en Colombia se obtienen al aplicar los modelos de regresión Cox PH y AFT para identificar factores en el cubo de información que tienen una incidencia estadísticamente significativa sobre la probabilidad de riesgo de las empresas. Estos factores se dividen según su incidencia en el riesgo; los factores Clase 1 son aquellos con incidencia directamente proporcional en el riesgo de desaparición, esto quiere decir que a mayor valor de esa variable, se estima mayor riesgo de desaparición de la empresa. Los factores Clase 2 son aquellos con una incidencia inversamente proporcional al riesgo de desaparición - a mayor valor de la variable, se estima un riesgo menor de desaparición. Los factores Clase 3 son aquellos para los cuales no se identifica una relación con el riesgo de desaparición. Al realizar este análisis utilizando dos métodos, se encuentran tanto coincidencias como diferencias entre la identificación de factores determinantes[7].

Entre los factores de **Clase 1**, se encuentra si la empresa recibió el **beneficio de la Ley 1780**, ley promulgada para impulsar la generación de empleo para jóvenes que otorga beneficios como exenciones en el pago de matrícula mercantil, o por ejemplo si la empresa es **"vendedor juego azar y suerte**" o el tipo de sociedad. Entre los factores **Clase 2** se encuentra **el tamaño de la empresa medido en activos (afectación inversa del 6%), la cantidad de establecimientos (23%) la actividad económica (18%), y el tipo de identificación de la empresa (26%)** entre otros.

[7] Para ver factores determinantes de la supervivencia, ver Anexo B. Vale recalcar que este anexo incluye las tablas con factores con significancia estadística, sin embargo y de acuerdo al consultor está "en lugar de confirmar una alta incidencia de factores en la probabilidad de supervivencia, permite identificar cuáles de ellos merecen un análisis posterior para verificar la verdadera magnitud de su aporte a las tasas de riesgo de desaparición"

Adicionalmente, este mismo análisis se realizó con los métodos de aprendizaje de máquinas, *Random Forest* y *Random Survival Forest*. Estos modelos arrojan que dieciséis factores[8] suman el 99% del puntaje máximo de importancia para determinar la supervivencia empresarial, entre los cuales se destacan también el **tamaño empresarial por activos, actividad económica principal, y los ingresos por actividad ordinaria**. Vale recalcar que aun cuando existen varias coincidencias entre los resultados de los modelos utilizados, los modelos de aprendizaje de máquina arrojan como importantes variables identificadas con relación 'nula' a través de modelos no-paramétricos, por ejemplo, "tamaño empresarial medido en activos".

En resumen, las variables identificadas como determinantes por los cuatros métodos se encuentra en la Tabla 3 y 4, abajo. La tabla 3 contiene factores determinantes y sus coeficientes y hazard ratios respectivos para Cox PH y AFT. La tabla 4, en comparación, incluye factores nuevos los cuales fueron desestimados por los métodos estadísticos pero cuya importancia es resaltada a través de los métodos de aprendizaje de máquinas.

Tabla 3. Factores determinantes de supervivencia empresarial (Cox PH y AFT)

Incidencia	Tipo Determinante	Determinante	Coeficiente (CoxPH)	Hazard Ratio[9] (Cox PH)	Coeficiente (AFT)[10]	Hazard Ratio (AFT)	P-value[11]	Método
Clase 1 Directa	Tipo de Empresa	Tipo sociedad	0.26	1.29	-0.57	0.56	<0.005	Cox/AFT/RF
		Organización jurídica	0.10	1.10	-0.01	0.99	<0.005	Cox/AFT/RF
		Beneficio Ley 1780	0.34	1.40	-0.64	0.53	<0.005	Cox/AFT/RF
	Ubicación	Municipio comercial	0.21	1.23	-0.65	0.52	<0.005	Cox/AFT/RF
		Cámara de Comercio	0.18	1.20	-0.52	0.60	<0.005	Cox/AFT/RF
	Internacionalización	Vocación Exportadora	0.77	2.15	-2.57	0.08	<0.005	Cox/AFT/RF
Clase 2 Inversa	Tipo de Empresa	Tamaño empleados	-0.02	0.98	-0.08	0.93	<0.005	Cox/AFT/RF
		Tamaño activos	-0.06	0.94	0.32	1.38	<0.005	Cox/AFT/RF
		CIIU principal	-0.20	0.82	0.65	1.91	<0.005	Cox/AFT/RF
		Cantidad	-0.26	0.77	1.18	3.26	<0.005	Cox/AFT/RF

[8] Entre estos factores se encuentra "último año renovado", "duración", "tamaño empresarial - activos", "CIIU principal", "ingresos actividad ordinaria", "cod. Municipio comercial", cod. Cámara de comercio", "CIIU secundario", "tamaño empresa por

[9] El *Hazard Ratio* es la suma de todas las tasas de riesgo posible que puede tomar cada factor. La interpretación de este valor se hace en función de un valor de 1 - por ejemplo, un *hazard ratio de* 2.15 quiere decir un incremento del 115% 0 1.15 veces mayor en la función de riesgo base

[10] A diferencia de Cox PH, para AFT los coeficientes son negativos para factores Clase 1 y positivos para factores Clase 2.

[11] La significancia en lugar de confirmar una alta incidencia, permite identificar factores que merecen análisis posterior para verificar su verdadera significancia.

		Establecimientos						
		Clase identificación	-0.30	0.74	0.29	1.34	<0.005	Cox/AFT/RF
	Indicadores Financieros	Ingresos actividad ordinaria	-0.09	0.91	0.31	1.36	<0.005	Cox/AFT/RF

Tabla 4. Factores determinantes de supervivencia empresarial (Random Forest y Random Survival Forest)

Tipo Determinante	Determinante	Importancia (RF)[12]	Importancia (RSF)	Método
Tipo de Empresa	CIIU secundario	0.02445	0.063	RF/RSF
	CIIU3	0.00799	0.028	RF/RSF
	Tamaño Empleados	0.02113	0.069	RF/RSF
	Organización jurídica	0.00819	0.067	RF/RSF
Indicadores Financieros	Capital social	0.00534	0.015	RF/RSF
	Capital Social Nacional Privado	0.00450	0.05200	RF/RSF
	Utilidad pérdida operacional	0.00714	0.05500	RF/RSF
	Resultado período	0.00556	0.03700	RF/RSF

Métodos utilizados: Cox PH, AFT, Random Forest, Random Survival Forest

3) ¿Cuál es la probabilidad de que una empresa empiece a exportar en Colombia?

Utilizando el método Kaplan Meier se puede determinar la probabilidad de la primera exportación de las empresas, necesitando solamente conocer la duración tras la consolidación de la empresa y su primera exportación. Según este modelo, se estima que **4 de cada 1000 empresas colombianas realizarán su primera exportación al término del primer año**. Esta figura se incrementa a **2 de cada 164 empresas al quinto año**, y 1 **de cada 65 empresas después de 10 años**.

Tabla 5. Probabilidad de Exportación de Empresas Colombianas

Año	Probabilidad de Supervivencia (%)
1	0.4
2	0.7

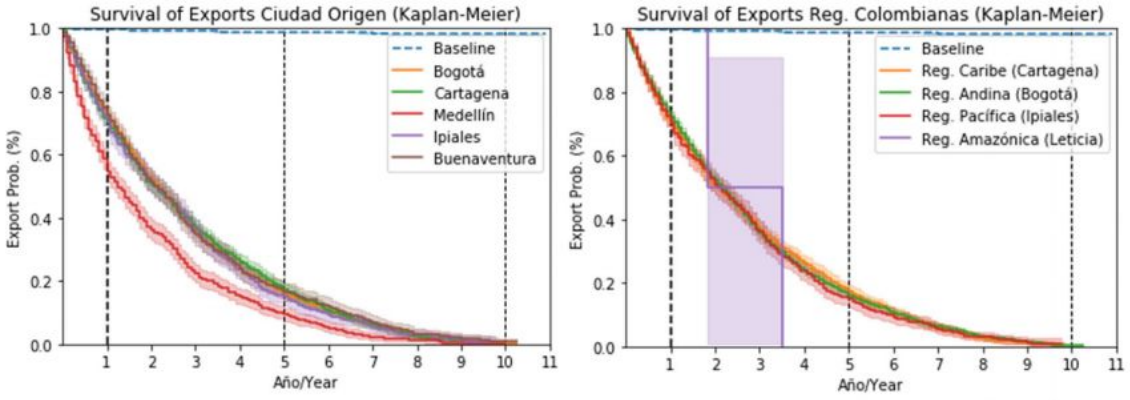
[12] La importancia revela la priorización que le da al modelo a los factores determinantes en la supervivencia. Estos valores son de 0 - 1, donde la suma total de las importancias equivale a uno.

3	0.9
5	1.2
10	1.6

Estas probabilidades coinciden con el hecho que de 5 millones de empresas, tan solo el 0.4% de estas exportan. Sin embargo, más allá de conocer las probabilidades de exportación por empresas, es interesante resaltar cómo estas cambian según el tiempo. Al ahondar el análisis y segmentar el grupo de empresas estudiadas por grupos de interés se encuentra que, por ejemplo, la probabilidad de que **empresas colombianas exporten hacia Sur o Norte América es más constante y estable que por ejemplo, a Europa y Asia**. Después del cuarto año, sin embargo, la **probabilidad de exportación hacia China** es mayor que el resto de países, con 81% de probabilidad de que esta sea su primera exportación. Además, se encuentra que quienes exportan “prendas y complementos (accesorios) de vestir, excepto los de punto” según el capítulo arancelario de las empresas, tienen una probabilidad de exportar por primera vez por encima de 30%, en comparación a los demás capítulos arancelarios que oscilan en un rango entre 15-25% de probabilidad.

Entre otros resultados, se encuentra también que **Medellín** es la ciudad que cuenta con mayor probabilidad de primera exportación a través de los años (44.8% a un año, 98.5% a 10 años). Similarmente, **Cartagena y Buenaventura,** quienes componen 18.5 y 9% respectivamente de las empresas exportadoras del país, obtienen una probabilidad de exportar alta, de 29.4% y 27% a un año respectivamente. Adicionalmente, el análisis arroja que el “Cod. Aduana Exp” (Código de la aduana desde la cual se exporta el producto), “Cod. Unidad Comercial Medida” (Nomenclatura de medida para la carga, e.g. Kg, Lt, cm cúbicos, etc.) y “Cod. País Destino” (País Destino) son factores adicionales determinantes.

Figura 2. Probabilidad de Exportación por Ciudad de Origen y por Regiones



4) ¿Cuáles son los factores determinantes para que una empresa se convierta en exportadora?

Similar al análisis realizado para determinar factores determinantes de supervivencia empresarial, se hace uso de Random Forest y Random Survival Forest para identificar factores determinantes en la primera exportación de una empresa colombiana.

Tabla 6. Factores determinantes de exportación (Random Forest y Random Survival Forest)

Factor	Fuente	Importancia (RF)	Importancia (RSF)
Cód. Aduana Exp	DIAN Expo	0.4020	0.1233
Cód. Unidad Comercial Medida	DIAN Expo	0.1926	0.0776
Cód. País Destino	DIAN Expo	0.1677	0.0471
Valor Fletes Posición	DIAN Expo	0.0751	0.0687
Duración		0.0374	0.0448
Cód. Aduana Impo	DIAN Impo	0.0177	0.0459
Otros Gastos	DIAN Impo	0.0164	0.0485
Valor Otros Gastos Posición	DIAN Expo	0.0152	0.0299
Activos	RUES	0.0088	0.0251
Ciudad Exportador	DIAN Expo	0.0066	0.0184
Cód. Unidad	DIAN Impo	0.0065	0.0365
CIIU Principal	RUES	0.0062	0.0154
CIIU Secundario	RUES	0.0054	0.0135
Utilidad Pérdida Operacional	RUES	0.0054	0.0480
Clasif. Importador Exportador	RUES	0.0050	0.0405
Empleados	RUES	0.0045	0.0327
CIIU3	RUES	0.0037	0.0061
Cód. Municipio Comercial	RUES	0.0035	0.0185
Ingresos Actividad Ordinaria	RUES	0.0031	0.0187
Cód Cámara	RUES	0.0028	0.0412
Último Año Renovado	RUES	0.0026	0.0471
CIIU4	RUES	0.0022	0.0125

Entre estos se resalta la importancia de la **ciudad donde se encuentra la empresa**, como un factor que presenta gran relevancia en determinar las probabilidades que tiene una empresa de iniciar actividades de comercio exterior, tanto actividades importadoras como exportadoras. Adicionalmente, el análisis arroja que el “Cod. Aduana Exp”, “Cod. Unidad Comercial Medida” y “Cod. Pais Destino” son factores determinantes adicionales.

IV. Retos y recomendaciones

Con el propósito de identificar oportunidades de mejoría de los prototipos de proyecto de analítica de datos adelantados por entidades del sector público, se hace una identificación de retos y obstáculos presentados a través de la realización del proyecto por parte de los consultores. Este proceso busca no solo identificar cuellos de botella, pero realizar sugerencias proactivas para mejorar la organización y desarrollo de proyectos de analítica de datos. Al ser este proyecto una de las primeras instancias donde entidades de orden nacional lideran proyectos de analitica de datos en colaboración con empresas privadas y talento local, cobra gran importancia la posibilidad de retroalimentar este proceso para así orientar al gobierno hacia un modelo generalizable y funcional de proyectos de analítica de datos.

La identificación de los retos y obstáculos se realizó identificando estos a lo largo de la cadena de valor de Big Data, o en general, en el proceso de uso de los datos recibidos en el marco de este proyecto. Estas etapas se resumen en la adquisición, almacenamiento, análisis, curaduría y uso de los datos.

Fuente: Freitas y Curry (2015) en “New Horizon for a Data-Driven Economy”

Los retos principales en la adquisición de los datos para este proyecto conciernen principalmente el acceso y tratamiento de los datos, así como el alcance del apoyo técnico brindado a los consultores. Principalmente, se resalta la falta de conocimiento de los datos por parte de iNNpulsa y el DNP. Dado que los datos utilizados en el proyecto provienen de fuentes externas a estas instituciones, se reconoce la ausencia de conocimiento detallado por parte de la rama técnica de estas entidades en el desenvolvimiento del proyecto. A pesar del importante rol, gestión y acompañamiento de iNNpulsa en asegurar el acceso a los datos y del apoyo a lo largo del proyecto, para proyectos futuros se recomienda contar con asesoramiento y apoyo técnico de las entidades dueñas y administradoras de los datos antes y durante el proyecto, quienes servirán de guía para acelerar el proceso y aclarar dudas frente a la recolección, definición de los datos, retos y obstáculos inherentes a las bases de datos utilizadas.

Particularmente, se identificó este obstáculo en el proceso de anonimización de los datos. Concretamente, para el inicio de este proyecto se realizó la anonimización con un identificador de contador, lo que dificulta la trazabilidad de las empresas a través de los años, característica crucial para los objetivos del proyecto. De igual manera, es importante recalcar que en el marco de este proyecto, hubo una ausencia de una plataforma o equipo experto en anonimización para facilitar este proceso.

Por otro lado, la necesidad de firmar convenios por separado entre las entidades para lograr acceder a los datos, presentó un obstáculo en el desarrollo del proyecto, al limitar los sujetos con posibilidad de manipular los datos. Esto presentó un reto logístico al limitar este manejo a iNNpulsa y únicamente en las instalaciones de esta entidad, el proceso de anonimización de los datos. Para proyectos futuros, recomendamos a las entidades administradoras de los datos generalmente, incluir convenios jurídicos más amplios para permitir con mayor facilidad el acceso y la reutilización de datos para estudios de esta naturaleza.

Entre otros cuellos de botella se identificaron retrasos en la entrega de los datos, y la ausencia de estandarización de las base de datos. Particularmente, se presentó un obstáculo en la obtención de diccionarios de datos para las fuentes compartidas, pues no todas las bases contaban con un diccionario existente, y para otras fuentes, lograr acceder a estos diccionarios fue un proceso complejo.

Por último, en términos de uso de los datos, un reto a futuro es el uso de los modelos objeto de este piloto. Se recomienda estructurar los términos de referencia de este tipo de proyectos no solo en busca de resultados concretos, pero también pensar en la reutilización de los modelos, algoritmos y script desarrollados. Una posibilidad a considerar es incluir entre los entregables productos fáciles de interpretar y utilizar para quienes no son expertos en la materia, por ejemplo dashboards o herramientas interactivas. También es relevante considerar necesidades futuras en el tema tratado, y considerar cómo estos proyectos pueden contribuir a la continuidad de análisis y construcción de habilidades. En otras palabras, diseñar fases piloto pensando en futuros proyectos y necesidades de las entidades involucradas, incluyendo gestión del conocimiento en los procedimientos de estos proyectos.

Recomendaciones Específicas: Proyecto Piloto de Supervivencia Empresarial

- Utilizar los productos de este piloto para retroalimentar la captura de los datos utilizados en el proyecto para entidades quienes administran y recolectan datos pertinentes a la supervivencia empresarial. Por ejemplo, el tamaño empresarial medido por cantidad de empleados y activos fue objeto de análisis tanto para la estimación de probabilidades de supervivencia, como para la identificación de los factores determinantes. No obstante, la naturaleza declarativa de estas variables impide la aceptación de los resultados con amplio margen de confianza. Por lo tanto, en futuros trabajos se recomienda la integración de fuentes de datos que soporten la validación de estas variables empresariales. Las fuentes propuestas son la Planilla Integrada de Liquidación de Aportes (PILA) para la cantidad de empleados y los reportes financieros de las Superintendencia de Sociedades y Superintendencia Financiera para la contrastación de activos y formas de capital.
- Para futuros trabajos e investigaciones de supervivencia empresarial, se recomienda incluir las empresas que hacen exportación de servicios, las cuales por reglamentación, únicamente hacen su registro inicial de exportaciones ante la DIAN y unas pocas solicitan apoyo en el registro como exportadora en Procolombia. Estos registros no son

información pública, pero su integración al presente modelo permitiría analizar todo el universo de exportaciones en Colombia y generar políticas públicas de apoyo a las empresas de servicios, las cuales están cada vez más interesadas en realizar planes de internacionalización.

- En el análisis de factores determinantes se debe tener en la cuenta la baja calidad de datos existente desde los conjuntos de datos originales, principalmente en términos de completitud. Por ejemplo, los factores: ingresos, actividad ordinaria, resultado del periodo, capital social, capital social nacional/extranjero público/privado, activos y empleados; cuentan con una gran mayoría de valores vacíos o en cero. En razón de esto, se recomiendan tomar acciones que lleven a la completitud o validación de estos factores desde la captura o re-ingeniería de los mismos.
- La búsqueda de factores determinantes relacionados con la educación de los fundadores y representantes legales de las empresas debe estar soportada por fuentes adicionales y de mayor impacto, tales como registros de acceso a educación superior, posgrados o formación académica en el exterior. La escolaridad en niveles de educación básica arrojó inicialmente un volumen de datos muy limitado en comparación con la cantidad de empresas, además que su calidad de datos se encuentra en niveles deficientes.
- En futuros trabajos se deben desarrollar o implementar mecanismos para identificar o aislar el efecto de las llamadas empresas de papel en las probabilidades de supervivencia y primera exportación. Para esto, se recomienda la integración de fuentes de datos de entidades con facultades de vigilancia y sanción, como la Superintendencia de Sociedades y la Superintendencia Financiera. Además, adoptar protocolos generados por estas entidades para la identificación o seguimiento de este tipo de sociedades ficticias.
- El valor agregado de las soluciones basadas en analítica predictiva no solo consiste en la capacidad de procesar grandes volúmenes de datos, sino también en la velocidad y recurrencia de tal procesamiento. Esto con el fin de obtener respuestas acerca de las probabilidades de supervivencia y primera exportación, cuando sean requeridas por iNNpulsa. Por lo tanto en trabajos futuros, se recomienda la implementación de un esquema periódico de despliegue y ejecución de los modelos desarrollados, de tal forma que se actualicen continuamente las respuestas de negocio, a partir de la nueva información disponible en las fuentes de datos base: RUES, DIAN y Mineducación.

Recomendaciones Generales: Proyectos de Analítica de Datos del Estado

- Establecer los estándares para la recepción de los datos en el marco de proyectos de datos en conjunto con los consultores. En lo posible, agendar tiempo antes del inicio del proyecto para reuniones entre administradores de datos y consultores.
- Promover el afianzamiento y la consolidación de las infraestructuras físicas adecuadas para adelantar proyectos de Big Data e Inteligencia Artificial en entidades

gubernamentales. Los recursos de procesamiento y almacenamiento resultan insuficientes para la preparación de los datos a ser entregados a los consultores y la validación de los resultados obtenidos. Lo anterior, redunda en conjuntos de datos incompletos, demoras prolongadas de integración y anonimización, reprocesos de preparación y sobretodo incapacidad para la replicación de resultados. Esto impone una gran limitante en el uso y aprovechamiento futuro de los modelos de analítica por parte de iNNpulsa de manera autónoma.

- Utilizar soluciones con mayores estándares de seguridad para la transferencia de datos. La recomendación principal es utilizar el SandBox de Big Data disponible para realizar transferencias de grandes volúmenes de datos entre diferentes entidades; esta no requiere esfuerzos de implementación y la confianza de los sistemas es alta.

- Priorizar y estandarizar la creación de diccionarios de datos para las fuentes de datos del Estado. El tener un diccionario exhaustivo y actualizado facilita la integración de diferentes fuentes, así como el entendimiento en general de los datos.

- Conformar equipos multidisciplinares con capital humano experto tanto en negocio como en datos para el uso y aprovechamiento de los modelos de análitica. Las entidades de Gobierno participantes de proyectos pilotos requieren equipos más robustos que puedan atender de manera especializada tanto los ámbitos de negocio como los técnicos sin llevar a afectaciones en el cumplimiento de hitos o tiempos de entrega del proyecto.

- Dentro de la cultura organizacional de datos se recomienda la creación de mecanismos e instrumentos que permitan la cooperación de las entidades del Gobierno para el intercambio efectivo y oportuno de fuentes de información. Los trabajos futuros que se puedan realizar en la evolución del modelo y el piloto, dependen en gran parte del acceso a nuevas fuentes de información bajo la custodia de otros organismos gubernamentales.

- Reforzar la implementación de guías gubernamentales en temas de interoperabilidad y calidad de datos (Guía Técnica de Interoperabilidad y Calidad de Datos de MinTIC, por ejemplo) para ahorrar esfuerzos en procesos de reingeniería y gestión de la arquitectura de datos de los proyectos. Verificar la calidad de los datos previo a la entrega de estos garantiza que de entrada, los datos utilizados sean fiables.

- Incluir en los entregables de estos proyectos procedimientos que construyan habilidades dentro de las organizaciones relevantes, incluyendo gestión del conocimiento y documental. Así mismo, priorizar entregables que sean de fácil uso para grupos de no-expertos. Se recomienda el desarrollo de una interfaz gráfica para la publicación y divulgación de los resultados de los modelos de analítica en lo que respecta a las probabilidades de supervivencia, primera exportación y factores determinantes. Por ejemplo, el uso de tableros de control o balanced scorecards facilitan el acceso a los resultados del ejercicio analítico para la toma de decisiones y sin mayores resistencias para las diferentes audiencias objetivo.

Anexo A. Variables - Cubo de Información con variables por agregación

#	Variable	Agreg.	#	Variable	Agreg.
1	Último Año Renovado	Max	2	Municipio fiscal	Moda
3	Importador/Exportador	Moda	4	Cód. Lugar Salida*	Moda
5	Vendedor Juego Azar Suerte	Moda	6	Cód. Régimen CAN*	Moda
7	Beneficio Ley 1780	Moda	8	Sistemas Especiales*	Moda
9	Cód. Tipo Sociedad	Moda	10	Cant. Unidades Pos.*	Media
11	Ciudad Exportador	Moda	12	Valor COP FOB Pos.*	Media
13	Cód. Municipio Comercial	Moda	14	País Compra*	Moda
15	Cód. Cámara de Comercio	Moda	16	Cód. Regimen*	Moda
17	Valor Ajuste	Media	18	Ctd. Unidades*	Media
19	Cód. País Destino	Moda	20	Valor CIF US Mercan.*	Media
21	Código Org. Jurídica	Moda	22	Ciudad Importador*	Moda
23	CIIU Secundario	Moda	24	Total IVA Otros Gastos*	Media
25	Actividad Económica	Moda	26	Dpto. Importador*	Moda
27	Cód. Unidad Comercial Medida	Moda	28	Cód. Categoría*	Moda
29	Capital Social Extranjero Público	Media	30	Dpto. Procedencia*	Moda
31	Capital Social Nacional Privado	Media	32	Cód. Modalidad*	Moda
33	Capital Social	Media	34	Posición Arancelaria Ex*	Moda
35	Otros Ingresos	Media	36	Total Kg Brutos*	Media
37	Capital Social Extranjero Privado	Media	38	Valor Seguro Posición*	Media
39	CIIU3	Moda	40	Dpto. Destino*	Moda
41	Utilidad Pérdida Operacional	Media	42	Cód. Acuerdo*	Moda
43	CIIU4	Moda	44	Posición Arancelaria Im*	Moda

45	Resultado Período	Media	46	Valor CIF COP Mercan.*	Media
47	Cód. Unidad*	Moda	48	Cód. Admon Aduana*	Moda
49	Derechos Arancelarios*	Media	50	Seguros*	Media
51	Cód. Aduana Exportación*	Moda	52	Cód. País Exportador*	Moda
53	Valor Fletes Posición*	Media	54	Cód. País Destino Num*	Moda
55	Empleados	Media	56	Cod. Vía Transporte Ex*	Moda
57	Capital Social Nacional Público	Media	58	Forma Pago*	Moda
59	Activos	Media	60	Dpto. Origen Posición*	Moda
61	Valor Otros Gastos Posición*	Media	62	Total Kg Netos*	Media
63	Porcentaje Arancel*	Media	64	País Origen*	Moda
65	Valor Agregado Nacional Posición*	Media	66	Cod. Vía Transporte Im*	Moda
67	Ingresos Actividad Ordinaria	Media	68	Peso Bruto Kgs*	Media
69	Cód. Aduana Importación*	Moda	70	Valor FOB US Mercan.*	Media
71	Otros Gastos*	Media	72	Impuestos Ventas*	Media
73	CIIU Principal	Moda	74	Valor Aduana*	Media
75	Cantidad Establecimientos	Media	76	Lugar Ingreso*	Moda
77	Cod Clase Identificación	Moda	78	Tipo Importación*	Moda
79	Cód. Lugar Salida Num*	Moda	80	Peso Neto Kgs*	Media
81	Nac. Medio Transporte*	Moda	82	Fletes*	Media
83	Tipo Cert. Origen*	Moda	84	Clase Importador*	Moda
85	Cód. Unidad Medida*	Moda	86	Base IVA*	Moda
87	Valor US FOB Posición*	Media	88	Cód. Lugar Ingreso*	Moda
89	País Procedencia*	Moda	90	Bandera*	Moda
91	NIT	NA	92	Razón Social	NA
93	Cód. Estado	NA	94	Fecha Matrícula	Min
95	Fecha Renovación	Max	96	Fecha Actualización Rues	Max

97	Fecha Cancelación	Max	98	Fecha Primera Expo*	Min
99	Fecha Última Expo*	Max	100	Dígito Verificación	NA
101	Otros Derechos*	Media	102	Porcentajes Otros*	Media
103	Base Otros*	Media	104	Tipo Cruce**	Media
105	Indíce Cruce**	Media	106	Cod Corte**	Media
107	Actividad**	Media	108	Nivel**	Media
109	Grado**	Media	110	Asiste**	Moda

Variables y agregaciones del cubo de información conformado. * Variables de DIAN Expo Impo. ** Variables de Mineducación

Las tablas abajo incluyen descripciones de la ocurrencia y porcentaje de los atributos de algunas de las categorías. Esto es para ilustrar el tipo de variables y la frecuencia de estas.

Tipo ID		
	Ocurrencia	Procentaje
CC	3644199	67.42
NIT	1740176	32.20
CE	13508	0.25
Pasaporte	5207	0.10
TI	1740	0.03
Doc Extranjería	32	0.00
Registro	22	0.00
Ninguna	1	0.00
Municipio		
Bogotá	1448793	26.81
Medellín	284400	5.26
Cali	274102	5.07
Barranquilla	181880	3.37
Bucaramanga	138093	2.55
Cúcuta	125900	2.33
Pereira	116323	2.15
Ibagué	110596	2.05

Org Jurídica		
Natural	4187804	77.48
SAS	640504	11.85
SA	381983	7.07
Unipersonales	68960	1.28
SA	45257	0.84
Expo/Impo		
Ninguna	5308357	98.21
Importador	59073	1.09
Mixta	16938	0.31
Exportador	13709	0.25
Empleados		
0	3916429	72.46
1	882192	16.32
2	257274	4.76
3	104090	1.93
4	58272	1.08
5	41924	0.78
6	23048	0.43

Cartagena	106839	1.98
Villavicencia	105552	1.95
Manizales	78765	1.46
Neiva	76384	1.41
Santa Marta	73179	1.35
Armenia	66420	1.23
Valledupar	52265	0.97
Pasto	48825	0.90
Popayán	45527	0.84
Monteria	42367	0.78
Itaguí	41413	0.77
Tunja	40777	0.75

10	16817	0.31
Activos		
0MM	2594609	48.00
1MM	308535	5.71
2MM	155794	2.88
1.5MM	122851	2.27
5M	113291	2.10
500K	100553	1.86
3M	97927	1.81

Anexo B - Fórmulas

Kaplan-Meier

La fórmula para estimar la probabilidad de supervivencia es:

$$\hat{S} = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$$

donde n_i es la cantidad empresas en riesgo de desaparición en un tiempo de interés t, d_i es la cantidad de empresas desaparecidas en un tiempo de interés t, y Π corresponde a la multiplicatoria de los operadores antes mencionados para todos los tiempos de interés t_i anteriores al tiempo límite t

Nelson-Aalen

El método estima la función de riesgo acumulada usando la fórmula:

$$\hat{H} = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

donde n_i es la cantidad de empresas en riesgo de desaparición en un tiempo de interés t, d_i es la cantidad de empresas desaparecidas en un tiempo de interés t.

Métodos Paramétricos (Distrubición Paramétrica Weibull)

Para estimar las probabilidades de supervivencia y tasas de riesgo sobre la totalidad de empresas bajo estudio, la fórmula utilizada es la siguiente: $\hat{S} = exp\left(-\left(\frac{t}{\lambda}\right)^{\rho}\right) \quad ; \quad \lambda > 0, \ \rho > 0$ donde λ, ρ son los parámetros de la distribución paramétrica Weibull, estimados a través del método de ajuste: *maximum likelihood estimation (MLE)*, el cual emplea una muestra del conjunto de datos suministrado para calcular los parámetros a través del método de optimización iterativa de Newton. El tiempo de convergencia y el número de iteraciones requeridas están en función del tamaño de la muestra de datos hasta satisfacer el criterio de parada (*stopping criterion*) Lipschitz-Hessian $\left\|x_{k+1} - x_*\right\| \leq 0.5\left\|x_k - x_*\right\|^2$, es decir, el error de la nueva iteración debe ser menor o igual al error cuadrado de la iteración anterior (Cameron 2019). El estudio en cuestión llevó a cabo una estimación exhaustiva, tomando la totalidad de las más de 5 millones de observaciones, en lugar de muestras parciales.

Estimados los parámetros λ, ρ, también se puede generar la función de riesgo acumulada, definida así $\hat{H}(t) = \left(\frac{t}{\lambda}\right)^{\rho}$ donde λ, ρ son los parámetros de la distribución paramétrica Weibull.

Cox's Proportional Hazard

Para estimar la función de riesgo global a partir de los aportes parciales de las covariantes, se utiliza la fórmula

$$h(t|x) = b_0(t) * exp\left(\sum_{i=1}^{n} b_i \cdot (x_i - \underline{x}_i)\right)$$

donde $b_0(t)$ es la línea base de riesgo, b_i es el coeficiente operador o peso de la variable o covariante, x_i es el covariante relacionado con la supervivencia empresarial. La expresión $b_i \cdot (x_i - \underline{x}_i)$ se denomina log de riesgo parcial y $exp\left(\sum_{i=1}^{n} b_i \cdot (x_i - \underline{x}_i)\right)$ corresponde al riesgo parcial, ya que estima el riesgo que aporta cada variable al riesgo global.

Tiempo de Falla Acumulado

Para estimar las funciones de supervivencia y riesgo, se utiliza la fórmula $\hat{S_A}(t) = S_B\left(\frac{t}{\lambda}\right)$; $\lambda(x) = exp\left(b_0 + \sum_{i=1}^{n} b_i \cdot x_i\right)$ donde λ es la tasa de falla acelerada dependiente los covariantes. Ahora, se debe seleccionar una función paramétrica, e.g. Weibull, cuya función de riesgo acumulada está también expresada en función de los covariantes. La función de riesgo acumulada es: $\hat{H}_B(t;x) = \left(\frac{t}{\lambda(x)}\right)^{\rho}$