# Evaluación de modelos de aprendizaje automático para la estimación del valor agregado en el sector manufacturero colombiano a partir de la EAM

Federico Hernández Guana

Asesor

Camilo Enrique Romero Parra

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

**Resumen**

El sector manufacturero es fundamental para la economía colombiana, por lo que comprender las dinámicas que impulsan su productividad es muy importante. Este proyecto aplicado implementa modelos de aprendizaje automático sobre los datos de la Encuesta Anual Manufacturera (EAM) del DANE, con el objetivo de desarrollar un modelo con alta capacidad de generalización para la estimación del valor agregado. El procesamiento de los datos abarcó un tratamiento de heterogeneidad mediante la eliminación de valores atípicos y el escalamiento de algunas variables, consolidando un conjunto de datos final con una división balanceada para ejecutar el entrenamiento y prueba. Se entrenaron y contrastaron modelos que fueron evaluados mediante métricas de desempeño. El algoritmo Random Forest fue seleccionado como el modelo ganador al alcanzar un resultado superior demostrando una alta robustez frente a la dispersión estructural de los datos. A través del análisis de importancia de características se determinó que la producción bruta y el consumo intermedio constituyen los predictores críticos de la variable objetivo. En conclusión, el modelo desarrollado provee una herramienta analítica con aplicación práctica para el análisis de consistencia de datos y el diseño de políticas orientadas a la optimización sectorial.

***Palabras claves:*** Valor agregado, aprendizaje automático, Random Forest, Encuesta Anual Manufacturera, métricas de evaluación.

**Abstract**

The manufacturing sector is fundamental to the Colombian economy, making it highly important to understand the dynamics that drive its productivity. This applied project implements machine learning models on data from DANE's Annual Manufacturing Survey (*Encuesta Anual Manufacturera* – EAM), aiming to develop a model with high generalization capacity for estimating value added. Data processing encompassed heterogeneity treatment through outlier removal and variable scaling, consolidating a final dataset with a balanced split to execute training and testing. Models were trained, contrasted and then evaluated using performance metrics. The Random Forest algorithm was selected as the winning model upon achieving superior results, demonstrating high robustness against the structural dispersion of the data. Through feature importance analysis, it was determined that gross production and intermediate consumption constitute the critical predictors of the target variable. In conclusion, the developed model provides an analytical tool with practical application for data consistency analysis and the design of policies oriented toward sectoral optimization.

***Keywords*:** Value added, machine learning, Random Forest, Annual Manufacturing Survey, evaluation metrics.

# Tabla de Contenido

Introducción ..... 9

Planteamiento del Problema ..... 11

Justificación ..... 14

Objetivos ..... 17

Objetivo General ..... 17

Objetivos Específicos ..... 17

Marco Teórico ..... 18

Esquema Conceptual ..... 26

Estado del Arte ..... 27

Metodología CRISP-DM ..... 29

Etapas de la Metodología CRISP-DM ..... 31

Comprensión del Negocio ..... 32

Comprensión de los Datos ..... 34

Preparación de los Datos ..... 35

Modelado ..... 36

Evaluación ..... 37

Despliegue ..... 39

Análisis de la Operación Estadística: Encuesta Anual Manufacturera (EAM) ..... 41

Características de la Operación Estadística ..... 41

Variables Principales ..... 42

Variables Específicas ..... 43

Análisis Exploratorio y Selección de Variables Críticas ..... 45

Obtención de los Datos ..... 45
Proceso de Recolección y Acopio (Metodología DANE) ..... 45
Recolección ..... 46
Acopio ..... 47
Diseño de Procesamiento ..... 48
Diseño del Análisis ..... 51
Descripción del Dataset ..... 54
Caracterización de las Variables ..... 55
Estadísticas Descriptivas ..... 57
Análisis de Heterogeneidad ..... 57
Análisis de Medidas de Posición ..... 58
Representatividad de la Muestra ..... 58
Diagnóstico por Variable ..... 58
Análisis de la Matriz de Correlación ..... 62
Diagnóstico de Viabilidad ..... 64
Procesamiento de los Datos ..... 66
Consolidación de la Base de Datos ..... 66
Preparación de los Datos ..... 67
Limpieza de Outliers y Valores Nulos ..... 68
Trazabilidad del Procesamiento de los Datos ..... 69
Imputación de Datos Faltantes ..... 70
Transformación y Escalamiento de Variables ..... 71
Descripción de la Base de Datos Final ..... 71

Entrenamiento del Modelo ..... 73
Comparación de Modelos ..... 73
Conformación del Conjunto de Entrenamiento y Prueba ..... 75
Entrenamiento y Optimización de Modelos ..... 75
Evaluación del Modelo ..... 77
Métricas Estadísticas ..... 77
Evaluación de las Métricas Estadísticas ..... 78
Selección del Modelo Ganador ..... 81
Análisis de Importancia de las Variables ..... 83
Generación de Combinaciones ..... 84
Conclusiones ..... 88
Recomendaciones ..... 91
Referencias Bibliográficas ..... 94

## Lista de Tablas

**Tabla 1** *Diagnóstico de Calidad de Datos* ..... 55

**Tabla 2** *Estadísticas Descriptivas* ..... 57

**Tabla 3** *Datos Consolidados* ..... 66

**Tabla 4** *Trazabilidad del Proceso de Limpieza* ..... 69

**Tabla 5** *Comparación de Modelos* ..... 78

**Tabla 6** *Importancia de Variables* ..... 83

## Lista de Figuras

**Figura 1** *Esquema Conceptual* ..... 26

**Figura 2** *Fases del Modelo CRISP-DM* ..... 32

**Figura 3** *Gráfico Boxplot de Variables Criticas* ..... 60

**Figura 4** *Matriz de Correlación (Variables Críticas)* ..... 62

**Figura 5** *$R^2$ Optimizado* ..... 86

## Introducción

El sector manufacturero colombiano cumple un papel estratégico en la economía nacional, al ser fuente generadora de empleo, innovación y desarrollo económico. Sin embargo, las empresas de este sector presentan diferencias en su desempeño económico y productivo, lo cual se ve reflejado en diferencias significativas en ingresos, productividad, inversión, contratación de personal y eficiencia en el uso de recursos.

La Encuesta Anual Manufacturera (EAM) es una herramienta estratégica para entender el comportamiento del sector industrial colombiano, pero la mayoría de los análisis sobre la industria manufacturera se han enfocado en análisis descriptivos o estadísticas convencionales, sin incorporar herramientas avanzadas de analítica predictiva que ayuden a anticipar el comportamiento de las empresas ni comprender de manera precisa los factores que determinan su desempeño. Actualmente no existe un análisis profundo que permita identificar los factores determinantes de los resultados obtenidos por los establecimientos industriales. Por ello, existe la necesidad de desarrollar modelos analíticos basados en datos que permitan identificar las variables que determinan el desempeño de los establecimientos y de esta forma generar insumos estratégicos para la toma de decisiones.

Este proyecto propone un análisis avanzado de los datos recolectados en la Encuesta Anual Manufacturera del DANE, empleando estadística descriptiva y técnicas de análisis de datos, para luego implementar técnicas de aprendizaje supervisado que permitan caracterizar el valor agregado generado por los establecimientos industriales de Colombia. A diferencia de los análisis descriptivos tradicionales, esta investigación busca transformar los datos históricos en una herramienta prospectiva que identifique los factores que realmente impulsan la generación de valor en la industria manufacturera.

Se selecciona la variable valor agregado como variable objetivo, ya que representa la verdadera contribución económica generada por los establecimientos industriales de Colombia. Con la implementación de un modelo de regresión sobre esta variable se busca identificar qué factores impulsan el crecimiento y así generar una herramienta esencial para la planificación industrial y la detección de ineficiencias operativas.

## Planteamiento del Problema

La industria manufacturera colombiana aporta cerca del 11% del PIB y es un importante generador de empleo, pero su competitividad se ve afectada por enfrentar retos como baja productividad, limitada innovación, desafíos tecnológicos, desaceleración económica y competencia (OECD, 2021). Respecto a las empresas manufactureras "se encuentra que su tasa de supervivencia a 5 años es del 33,4%, mientras que la tasa de supervivencia de empresas pequeñas es del 60,9%, la de empresas medianas del 73,7% y la de unidades grandes del 85,7%." (Confecámaras, 2023, p. 7).

Diversos estudios han abordado el crecimiento empresarial en Colombia. En particular, se hallaron estudios que analizan los determinantes del crecimiento empresarial del sector manufacturero en Colombia, con particular énfasis en factores internos como el capital humano, el valor agregado, el valor de las ventas, los costos de producción, administración y ventas (Carmona et al., 2020).

Por otro lado, existen investigaciones que han analizado los factores que contribuyen al cierre temprano de empresas, destacando la falta de control interno, acceso limitado a financiamiento y baja capacidad de adaptación al entorno competitivo (Silva, 2025).

Si bien la literatura sobre la industria manufacturera en Colombia es amplia en términos descriptivos, existe un vacío en la implementación de modelos de aprendizaje supervisado aplicados a los datos de la EAM para el modelado y caracterización de la variable valor agregado, lo que limita la capacidad de realizar análisis prospectivos sobre la eficiencia operativa del sector manufacturero.

Aunque en los últimos años se ha incrementado el uso de técnicas de aprendizaje automático en el ámbito empresarial y financiero, la literatura aún muestra una presencia

importante de métodos estadísticos y econométricos tradicionales para el análisis y la predicción de datos (Gao et al., 2024). Diversas revisiones destacan que muchos estudios continúan empleando modelos clásicos como referencia metodológica, mientras que los enfoques de *machine learning* se integran poco a poco como herramientas complementarias o comparativas (Fierro et al., 2022). Esta tendencia evidencia una transición gradual desde enfoques tradicionales hacia metodologías más avanzadas capaces de explicar patrones complejos en grandes volúmenes de datos.

La Encuesta Anual Manufacturera (EAM) se enfoca en los establecimientos colombianos que tienen como actividad principal la manufacturera. Esta se define como la transformación física o química de los materiales, sustancias y componentes, en productos nuevos; cuyo trabajo se puede realizar con máquinas o a mano, y en una fábrica o a domicilio. Además, esta investigación se enfoca en una muestra de establecimientos industriales, que proporcionan información detallada de variables como producción, empleo, costos, inversión y valor agregado. Sin embargo, su uso ha estado orientado principalmente a análisis descriptivos, sin aprovechar su potencial para desarrollar modelos basados en ciencia de datos. Actualmente, no se analizan adecuadamente las empresas industriales para poder entender su desempeño y las variables que impactan en sus resultados. Por tanto, surge la necesidad de desarrollar un modelo de analítica de datos y de aprendizaje supervisado que utilice la información histórica de la EAM para entender las variables que impactan el valor agregado y de esta forma generar un insumo para la toma de decisiones (Departamento Administrativo Nacional de Estadística (DANE), 2025c).

El problema central no radica únicamente en la falta de información, sino en la ausencia de modelos que permitan comprender el valor agregado generado por los establecimientos industriales. Sin esta capacidad, las empresas y las instituciones reguladoras operan bajo un

enfoque reactivo, limitando la posibilidad de intervenir oportunamente en sectores donde la productividad muestra signos de estancamiento.

Con base en el problema identificado se plantea la siguiente pregunta de investigación: ¿Cómo pueden los algoritmos de aprendizaje supervisado predecir con precisión el valor agregado de los establecimientos industriales en Colombia, utilizando las variables de producción, consumo intermedio, personal permanente y activos fijos reportadas en la Encuesta Anual Manufacturera?

## Justificación

El sector manufacturero constituye uno de los pilares del desarrollo productivo de Colombia, dado su impacto en la generación de empleo formal, la innovación y el crecimiento económico. Por esta razón, comprender los factores que inciden en el desempeño de los establecimientos manufactureros es fundamental para fortalecer la competitividad industrial y orientar políticas públicas basadas en evidencia. En este contexto, la disponibilidad de microdatos de la Encuesta Anual Manufacturera del DANE permite desarrollar modelos analíticos que faciliten la identificación de variables esenciales en el desempeño de los establecimientos industriales estudiados.

Diversas investigaciones han mostrado que el análisis predictivo empresarial permite tomar decisiones estratégicas oportunas, mejorar la asignación de recursos y diseñar programas de apoyo empresarial más efectivos (Tascón & Castaño, 2012). De igual forma, la literatura señala que combinar análisis exploratorio y modelos inferenciales permite descubrir patrones ocultos y validar relaciones significativas entre variables económicas, lo cual es clave para la construcción de conocimiento riguroso (Huber et al., 2017).

Adicionalmente, el auge de las técnicas de análisis estadístico y de machine learning ha demostrado ser especialmente útil en contextos complejos con múltiples variables económicas e interacciones no lineales (Buitrago Vargas, 2021). Estas metodologías permiten comparar modelos tradicionales como la regresión logística con enfoques más avanzados (bosques aleatorios o máquinas de soporte vectorial) para lograr mayor precisión predictiva sin perder la interpretación relevante para las personas que toman las decisiones (López et al., 2025).

Este proyecto aplicado implementa modelos estadísticos y de aprendizaje automático sobre los datos de la Encuesta Anual Manufacturera con el fin de identificar las variables

determinantes de la productividad industrial. A través de la caracterización de la variable valor agregado, el estudio permite modelar patrones de desempeño empresarial y consolidar un modelo predictivo capaz de generar estimaciones precisas que sirvan como insumo estratégico para la toma de decisiones.

Esto facilita la toma de decisiones informadas, el diseño de políticas focalizadas y la asignación de recursos de manera eficiente. Al caracterizar las empresas según sus variables, se podrán identificar los factores que presentan debilidad y diseñar estrategias necesarias para su fortalecimiento. Esta capacidad de entendimiento es clave para evitar cierres y fomentar la sostenibilidad empresarial. Además, este análisis contribuye a cerrar los vacíos en el uso de datos para una toma de decisiones más informada y a fortalecer las competencias analíticas de las personas encargadas de tomar decisiones.

Este proyecto aportará herramientas para la optimización de políticas industriales y estrategias empresariales, ofreciendo una base técnica para comprender mejor la dinámica del sector y fortalecer la gestión empresarial basada en evidencia cuantitativa. Para ejecutar este proyecto se realizará un análisis exploratorio de datos, ya que algunos autores señalan (Pramanik et al., 2019) que el análisis exploratorio respaldado por lenguajes como Python proporciona una estructura sólida para preparar y visualizar datos, mientras que modelos predictivos como los utilizados en scikit-forecasts facilitan la proyección de fenómenos económicos complejos (Arango, 2021).

Si bien existen estudios sobre competitividad, implementación de estrategias y productividad en el sector manufacturero colombiano, la mayoría de estos se han limitado a enfoques de carácter descriptivo. En este sentido, se evidencia un vacío en la aplicación de modelos predictivos basados en técnicas de ciencia de datos, particularmente aquellos que

empleen de manera directa la información recolectada por la Encuesta Anual Manufacturera del DANE. En atención a esta limitación, la presente investigación propone la utilización de modelos de regresión orientados a capturar las complejidades estructurales de la industria manufacturera colombiana, con el propósito de pasar de un enfoque descriptivo hacia uno de carácter prospectivo. Para tal fin, se selecciona el valor agregado como variable objetivo, ya que constituye un indicador representativo de la eficiencia productiva.

# Objetivos

## Objetivo General

Evaluar el desempeño de diferentes modelos de aprendizaje automático para la estimación del valor agregado en los establecimientos de la Encuesta Anual Manufacturera (EAM), con el fin de identificar el algoritmo con mayor capacidad predictiva y determinar la importancia relativa de las variables que impulsan la productividad industrial.

## Objetivos Específicos

Realizar un análisis exploratorio de los datos de la Encuesta Anual Manufacturera, mediante la generación de estadística descriptiva y análisis de correlación, para identificar y caracterizar las variables críticas que impactan el comportamiento productivo de la industria manufacturera.

Procesar los datos de la Encuesta Anual Manufacturera mediante técnicas de limpieza, tratamiento de valores atípicos y normalización de variables, con el fin de conformar un dataset consistente que garantice la robustez y convergencia de los algoritmos de aprendizaje supervisado.

Comparar diversos algoritmos de aprendizaje supervisado mediante el entrenamiento y validación de modelos lineales y no lineales, con el fin de seleccionar el modelo con mayor capacidad de generalización para la comprensión del valor agregado en el sector manufacturero.

Evaluar el desempeño de los modelos de aprendizaje supervisado mediante el cálculo de métricas de evaluación para modelos de regresión ($R^2$, MAE, RMSE y MAPE), para validar la utilidad de las estimaciones en la toma de decisiones y comprender los factores determinantes de la generación de valor agregado en el sector manufacturero.

## Marco Teórico

El presente capítulo define el marco teórico con el cual se busca contextualizar el problema de investigación enfocado en el desempeño empresarial en el sector manufacturero colombiano basado en la información de la Encuesta Anual Manufacturera (EAM), explicando los conceptos clave involucrados que sustenten la aplicación de técnicas de ciencia de datos para la caracterización de los establecimientos estudiados en la operación estadística.

La Encuesta Anual Manufacturera (EAM) es una operación estadística del DANE que recoge información detallada sobre los establecimientos manufactureros en Colombia. Incluye variables como producción, empleo, remuneraciones, energía consumida, consumo intermedio y valor agregado. El valor agregado es el total de los ingresos recibidos por el uso de los factores productivos (tierra, capital, trabajo) en el proceso de producción durante el período estudiado. El DANE calcula el valor agregado descontando del valor de la producción bruta el valor del consumo intermedio (Departamento Administrativo Nacional de Estadística (DANE), 2025a).

El análisis del desempeño empresarial busca evaluar la capacidad de una organización para generar valor económico y mantenerse competitiva en el mercado. Según Tascón y Castaño (Tascón & Castaño, 2012) la literatura internacional ha pasado de enfoques centrados exclusivamente en la identificación de empresas en quiebra hacia modelos predictivos capaces de anticipar la evolución del desempeño empresarial, empleando indicadores financieros y no financieros.

El análisis exploratorio de datos constituye una fase esencial en estudios cuantitativos, dado que permite examinar la estructura interna de los datos, identificar patrones, valores atípicos y relaciones preliminares entre variables antes de aplicar modelos confirmatorios. El proceso investigativo requiere combinar un enfoque exploratorio para descubrir estructuras no

previstas con metodologías confirmatorias que permitan validar hipótesis planteadas (Parra, 2002). Asimismo, el análisis exploratorio de datos contribuye a la comprensión profunda del comportamiento de los datos tanto en escenarios cualitativos como cuantitativos, facilitando el desarrollo de modelos explicativos y predictivos sólidos; permitirá analizar tendencias productivas, niveles de participación laboral, inversión y productividad en las empresas manufactureras colombianas (Huber et al., 2017).

Los métodos de aprendizaje estadístico permiten modelar patrones a partir de los datos mediante algoritmos que aprenden relaciones entre variables. Estos métodos se dividen en dos enfoques principales: Modelos supervisados y no supervisados. Los modelos supervisados buscan predecir una variable objetivo a partir de variables independientes, mientras que los no supervisados permiten identificar estructuras latentes en los datos. Los métodos de aprendizaje estadístico permiten modelar relaciones complejas entre variables, identificar patrones no visibles a simple vista y predecir comportamientos futuros (Buitrago, 2021).

Para implementar este modelo de aprendizaje es necesario ejecutar acciones de analítica de datos, la cual corresponde al proceso de análisis computacional sistemático de los datos del modelo relacional. El objetivo de esta etapa es descubrir, visualizar, interpretar y comunicar patrones significativos o tendencias generales en los datos (Araque & Giampietro, 2023).

Una de las herramientas esenciales para la implementación de este modelo es el aprendizaje automático (Machine Learning) el cual es un campo de estudio que da a las computadoras la capacidad de aprender sin ser explícitamente programadas. El aprendizaje se produce cuando un sistema mejora su rendimiento en una tarea específica a partir de la experiencia. En lugar de codificar reglas manualmente, se entrenan modelos que generalizan patrones y realizan predicciones. Machine Learning es una herramienta poderosa para tareas

como clasificación, regresión, detección de anomalías, agrupamiento y reducción de dimensionalidad. El aprendizaje automático es un conjunto de métodos estadísticos y computacionales que permiten a las máquinas aprender de la experiencia (datos) y mejorar su desempeño en tareas específicas sin necesidad de instrucciones rígidas (Gerón, 2019).

Otra metodología esencial para el desarrollo de este proyecto es la minería de datos la cual es una técnica cuyo propósito es extraer información valiosa y útil para la toma de decisiones, a partir del análisis y descubrimiento de patrones, tendencias y relaciones en grandes conjuntos de datos. En términos generales, la minería de datos implica la aplicación de algoritmos y técnicas estadísticas y computacionales avanzadas para explorar, analizar y modelar grandes conjuntos de datos (Ahumada, 2023).

Un modelo de regresión se define como una herramienta estadística utilizada para encontrar asociaciones de causa y efecto entre variables independientes y dependientes, siendo los tipos comunes la regresión lineal simple, la regresión lineal múltiple y las regresiones no lineales como la logística y la polinómica (Hazelton, 2010).

La estadística descriptiva es la rama de la estadística que se encarga de organizar, resumir y presentar los datos de manera clara y significativa, utilizando tablas, gráficos y medidas numéricas como promedios, variación y distribución. Su propósito es describir el comportamiento de un conjunto de datos sin inferir conclusiones más allá de la muestra analizada (Triola, 2004.). En este proyecto es esencial para realizar un análisis inicial que defina la estructura y comportamiento de las variables de la Encuesta Anual Manufacturera.

Otra herramienta esencial es Scikit-learn, la cual es una librería simple y eficiente de aprendizaje automático, enfocada en herramientas para la predicción y el análisis de datos. A diferencia de las demás librerías, Scikit-learn cuenta con diferentes funcionalidades específicas

para el análisis de series de tiempo. Una de ellas es la capacidad de diferenciar los modelos entre clasificación y regresión (Arango, 2021).

Los algoritmos de ensamble se describen como métodos que combinan múltiples modelos (por ejemplo, varios árboles de decisión) para producir un predictor más robusto y preciso que cualquiera de los modelos individuales. La idea central es que, en lugar de depender de un único modelo, se construye un conjunto de modelos cuyas predicciones se agregan (por votación en clasificación o promediando en regresión). Esto reduce la varianza y mejora la generalización, ya que los errores de cada modelo tienden a compensarse entre sí. Un algoritmo de ensamble es una estrategia para mejorar el rendimiento de los modelos de aprendizaje automático mediante la combinación de varios predictores en lugar de confiar en uno solo (James et al., 2023).

Respecto al procesamiento de los datos se seleccionó la herramienta Python ya que se ha consolidado como un instrumento clave en ciencia de datos por su versatilidad, accesibilidad y amplio ecosistema de librerías. Esta herramienta se caracteriza por su utilidad para ejecutar EDA, limpieza de datos y visualizaciones mediante bibliotecas como Pandas, Matplotlib y Seaborn (Pramanik et al., 2019).

Un árbol de decisión es un algoritmo de aprendizaje supervisado no paramétrico que divide el espacio de los datos de forma recursiva en subconjuntos cada vez más homogéneos. En tareas de regresión, el algoritmo selecciona en cada nodo la variable y el punto de corte que minimizan la varianza o el error cuadrático medio de la variable objetivo. El proceso finaliza en hojas, donde la predicción final es el promedio de los valores de las observaciones que alcanzaron dicha terminal. Aunque son altamente interpretables, los árboles individuales tienden

al sobreajuste, lo que justifica el uso de métodos de ensamble para mejorar la estabilidad de la estimación (James et al., 2023).

Un árbol de decisión para regresión se define como un modelo predictivo que divide el espacio de los predictores en regiones más pequeñas y homogéneas, con el objetivo de aproximar una variable de respuesta continua. El procedimiento consiste en realizar particiones sucesivas de los datos, en cada nodo se selecciona una variable explicativa y un punto de corte que minimice la variabilidad dentro de cada región. En lugar de asignar una clase (como en clasificación), cada región terminal del árbol devuelve un valor numérico, que suele ser el promedio de la variable respuesta en esa región. Un árbol de regresión predice valores continuos dividiendo los datos en grupos cada vez más homogéneos y asignando a cada grupo un valor promedio (James et al., 2023).

También se seleccionó el algoritmo Random Forest el cual es un algoritmo de aprendizaje supervisado basado en el ensamble de múltiples árboles de decisión, diseñado para mejorar la precisión predictiva y reducir la varianza del modelo. El Random Forest introduce un proceso de correlación entre los árboles al seleccionar, en cada división de nodo, un subconjunto aleatorio de predictores del total de variables disponibles. El funcionamiento interno del algoritmo puede deducirse de su nombre: *random* (porque introduce una capa de aleatoriedad en cada modelo que se construye) y *forest* (porque se construyen varios modelos de árboles) (Ramasubramanian & Moolayil, 2019). Este mecanismo evita que los predictores dominantes (aquellos con mayor capacidad explicativa inicial) sesguen la totalidad del bosque, permitiendo que variables con efectos moderados participen en la construcción del modelo. El resultado final es un promedio de las predicciones de todos los árboles individuales, lo que genera una estructura robusta frente al sobreajuste (*overfitting*) y altamente eficiente para gestionar la

heterocedasticidad y las relaciones no lineales presentes en los datos industriales (James et al., 2023).

Para validar la utilidad del modelo frente a los datos analizados se utilizaron las siguientes métricas (Ramasubramanian & Moolayil, 2019):

- Coeficiente de determinación ($R^2$): Seleccionado para medir la proporción de la varianza del valor agregado que el modelo analizado logra capturar. Indica qué porcentaje de la variabilidad del valor agregado es explicada por el modelo. Un valor cercano a 1 indica una precisión alta.
- MAE (Error Medio Absoluto): Representa el promedio de cuánto se equivoca el modelo, se expresa en la misma escala que los datos originales.
- RMSE (Root Mean Squared Error): Es la raíz del error cuadrático medio e indica fallos grandes, por lo cual permite entender que tan mal le va al modelo especialmente con grandes errores.
- MAPE (Error Porcentual Absoluto Medio): Mide el tamaño del error promedio en relación con los valores reales. En lugar de decir en cuántas unidades se falló (como el MAE), muestra que tan grande fue ese fallo respecto al tamaño de la empresa.

Los algoritmos de ensamble (ensemble learning) se definen como métodos que combinan múltiples modelos para mejorar la precisión y la estabilidad de las predicciones. En lugar de depender de un único modelo, se construye un conjunto de modelos (por ejemplo, varios árboles de decisión). Las predicciones de estos modelos se agregan, ya sea mediante votación (en clasificación) o promediando (en regresión). Esto ayuda a reducir la varianza y el riesgo de sobreajuste, ya que los errores individuales tienden a compensarse. En resumen, un algoritmo de

ensamble es una estrategia para aprovechar la fuerza colectiva de varios modelos, logrando predicciones más robustas que las de un modelo individual (Hastie et al., 2017).

El aprendizaje supervisado se define como el método con el que se busca predecir el valor de una variable de salida a partir de un conjunto de variables de entrada. Se llama supervisado porque el proceso de entrenamiento está guiado por la presencia de la variable respuesta en los datos. El objetivo es construir un modelo que, a partir de ejemplos con entradas y salidas conocidas, pueda generalizar y hacer predicciones sobre nuevas observaciones. Mientras que el aprendizaje no supervisado se caracteriza porque no existe una variable de salida y la tarea consiste en descubrir patrones o estructuras en los datos. En resumen, el aprendizaje supervisado es el marco en el que se entrena un modelo con pares de entrada-salida conocidos, para que luego pueda predecir la salida de nuevas entradas (Hastie et al., 2017).

El *overfitting* ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento, capturando no solo las tendencias reales sino también el ruido o las fluctuaciones accidentales. Esto provoca que el modelo tenga un desempeño excelente en los datos usados para entrenarlo, pero falle al generalizar en datos nuevos (Massaron & Boschetti, 2016).

El *scaling* consiste en transformar las variables numéricas para que estén en una misma escala o rango. Esto es importante porque los algoritmos de regresión y aprendizaje automático suelen ser sensibles a la magnitud de las variables ya que una variable con valores muy grandes puede dominar el cálculo de los coeficientes y afectar la interpretación (Massaron & Boschetti, 2016).

El *gridsearch* es un procedimiento sistemático para encontrar los mejores hiperparámetros de un modelo. Se construye una rejilla (grid) con todas las combinaciones posibles de valores de los parámetros que se quieren ajustar. El algoritmo entrena y evalúa el

modelo con cada combinación, normalmente usando validación cruzada y selecciona la que produce el mejor desempeño. Este procedimiento explora exhaustivamente todas las combinaciones de parámetros en un espacio definido y se utiliza junto con técnicas de validación (como *cross-validation*) para evitar el sobreajuste (Massaron & Boschetti, 2016).

El tratamiento de outliers (Valores atípicos) se aborda dentro de la estadística descriptiva. Un outlier se define como una observación que se encuentra considerablemente alejada del resto de los datos, es decir, un valor extremo que no sigue el patrón general de la distribución. Estos valores pueden aparecer por las siguientes razones:

- Errores de medición o registro.
- Variabilidad natural del fenómeno.
- Condiciones especiales en el experimento.

El tratamiento de outliers implica decidir si se eliminan del análisis (cuando se sospecha que son errores), se mantienen (cuando representan variabilidad real) o se ajustan mediante transformaciones estadísticas o técnicas robustas que reduzcan su influencia en medidas como la media y la desviación estándar. Los outliers deben ser identificados y analizados cuidadosamente, ya que pueden distorsionar las conclusiones estadísticas si no se manejan de manera adecuada (Devore, 2018).

En síntesis, el marco teórico en este capítulo consolida una visión integral que articula la teoría de la producción industrial con las temáticas tecnológicas del análisis de datos. La revisión exhaustiva de las variables de la Encuesta Anual Manufacturera (EAM) genera la base del proyecto aplicado necesaria para entender qué mide el valor agregado, mientras que el estudio de los algoritmos de aprendizaje supervisado proporciona las herramientas requeridas para su estimación. De este modo, los conceptos aquí expuestos dejan de ser nociones abstractas y se

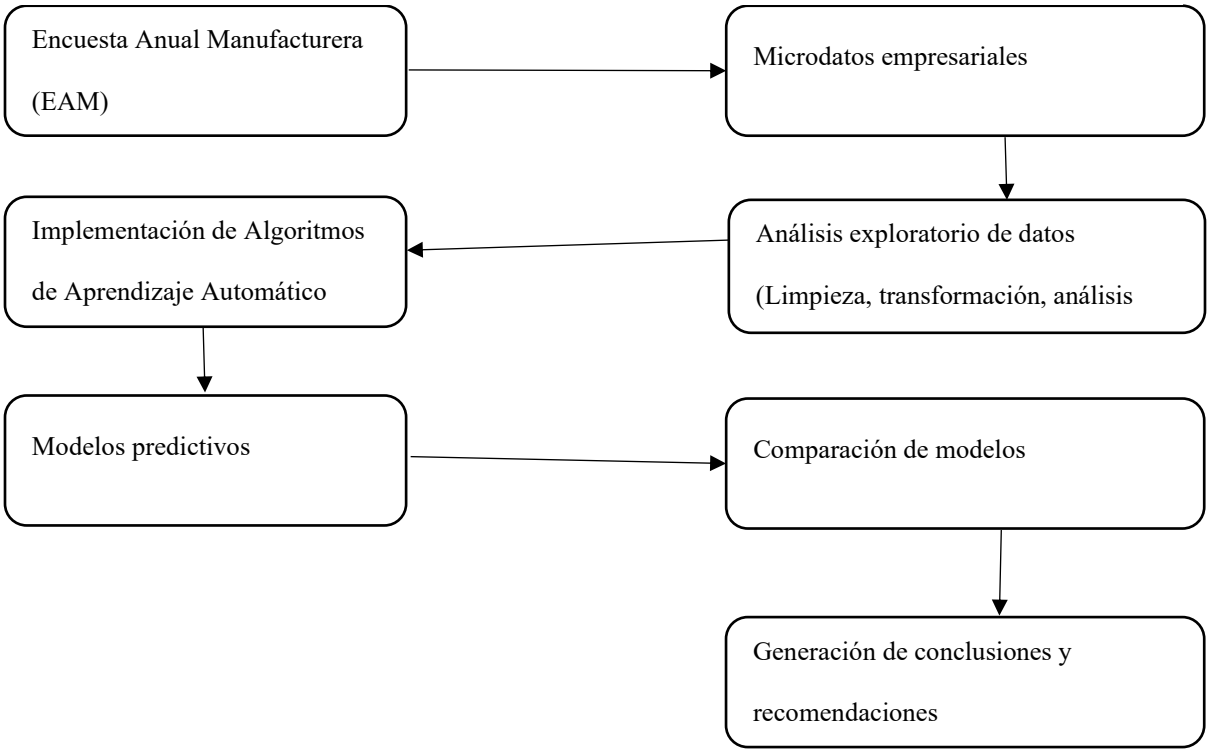
transforman en los pilares fundamentales que guiarán el diseño de este proyecto aplicado. Con este sustento teórico establecido, más adelante se detallarán las fases de la metodología CRISP-DM implementadas para la preparación de los datos y el entrenamiento de los modelos predictivos.

**Esquema Conceptual**

A continuación, se presenta un gráfico que permite entender el proceso de este proyecto aplicado.

**Figura 1**

*Esquema Conceptual*




*Nota*. Proceso desarrollado en el proyecto aplicado.

## Estado del Arte

En la última década la literatura especializada ha abordado el crecimiento empresarial en Colombia, diferentes investigaciones (Carmona González et al., 2020) destacan que el desempeño del sector manufacturero no es un fenómeno aislado, sino que responde a una interacción compleja de factores como el capital humano, el valor de las ventas, los costos de producción, administración y ventas. No obstante, se encontró que muchos estudios continúan empleando modelos clásicos, mientras que los enfoques de machine learning se integran poco a poco como herramientas complementarias o comparativas (Fierro Torres et al., 2022).

Por otro lado, existen investigaciones que se han enfocado en analizar los factores que contribuyen al cierre temprano de empresas, destacando la falta de control interno robusto, acceso limitado a mecanismos de financiación y baja capacidad de adaptación al entorno competitivo. Estas investigaciones exponen que el éxito no solo depende de la acumulación de capital, sino de la capacidad de adaptación al entorno competitivo (Silva, 2025). Aunque en los últimos años se ha incrementado el uso de técnicas de aprendizaje automático en el ámbito empresarial y financiero, la literatura aún muestra una presencia importante de métodos estadísticos y econométricos tradicionales para el análisis y la predicción de datos (Gao et al., 2024).

A partir de los análisis realizados se halló un vacío en la literatura sobre modelos de aprendizaje automático aplicados a los datos de la Encuesta Anual Manufacturera que permitan modelar y caracterizar el valor agregado. La mayoría de los estudios existentes se limitan a análisis descriptivos, careciendo de enfoques de regresión avanzada que faciliten la comprensión de los factores que impactan en la generación de valor. A pesar de la riqueza de los datos de la Encuesta Anual Manufacturera, la aplicación de algoritmos de aprendizaje automático sigue

siendo un área poco explorada, lo que lleva a que a la información obtenida en la EAM se le haya dado un uso orientado principalmente a análisis descriptivos. Actualmente no se aprovechan las herramientas avanzadas para desarrollar modelos basados en ciencia de datos que permitan entender el comportamiento de las empresas, los factores que determinan su desempeño o el impacto de sus principales variables. Por tanto, existe la necesidad de desarrollar un modelo de aprendizaje supervisado que utilice la información de la EAM para caracterizar el valor agregado e identificar el impacto de las variables críticas como las ventas, el empleo y los activos fijos en el resultado económico de las empresas manufactureras.

En conclusión, la revisión de la literatura evidencia que, si bien la Encuesta Anual Manufacturera es el pilar de la información industrial en Colombia, su aprovechamiento analítico se ha mantenido principalmente en un plano descriptivo y estructural. El vacío identificado radica en la falta de herramientas que gestionen la multidimensionalidad propia de estos datos sin sacrificar la interpretabilidad. Por consiguiente, este proyecto aplicado propone la implementación de modelos de aprendizaje automático aplicados a los datos de la EAM con el fin de caracterizar la variable valor agregado. Este enfoque permite llenar los vacíos analíticos tradicionales al modelar relaciones no lineales complejas, generando así insumos estratégicos que facilitan la toma de decisiones basadas en evidencia estadística.

## Metodología CRISP-DM

El presente proyecto aplicado adopta la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), reconocida como el estándar más utilizado para el desarrollo sistemático de proyectos de minería de datos y analítica predictiva. Esta metodología establece una estructura iterativa de 6 fases permitiendo el desarrollo riguroso, reproducible y orientado a resultados en proyectos de análisis de datos (Schröer et al., 2021). La metodología CRISP-DM facilita un proceso flexible y adaptable que guía desde la comprensión del problema hasta la ejecución y evaluación de modelos predictivos, promoviendo un vínculo entre objetivos organizacionales y técnicas analíticas avanzadas, para de esta forma hacer que los grandes proyectos de minería de datos sean menos costosos, más confiables, más repetibles, más manejables y rápidos (Wirth & Hipp, 2000).

CRISP-DM es una metodología caracterizada porque los datos están en el centro de todas las actividades de ciencias de datos. En el momento de implementar este proceso no siempre se deben seguir sus etapas de manera lineal, dependiendo de una etapa particular, se puede volver a una de las etapas anteriores, rehacer la etapa actual o pasar a la etapa siguiente (Kelleher & Tierny, 2022). Su carácter iterativo, semiestructurado y flexible la hace útil para proyectos aplicados, donde se requiere retroalimentación constante y se tienen los datos en el centro del proceso. CRISP-DM es una metodología que se puede describir mediante un modelo de proceso jerárquico, que consiste en tareas descritas en cuatro niveles de abstracción (de lo general a lo específico): Fase, tarea genérica, tarea especializada e instancia del proceso (Chapman et al., 2000).

La elección de CRISP-DM se justifica por su capacidad para asegurar que cada fase esté alineada con los objetivos del proyecto y del análisis técnico ya que permite abordar

sistemáticamente desde la comprensión del problema y la preparación de los datos hasta la evaluación e implementación del modelo predictivo. Este enfoque permite integrar técnicas de análisis exploratorio, modelado estadístico y aprendizaje automático, facilitando la interpretación de resultados y su aplicación en decisiones estratégicas.

En la guía de la metodología CRISP-DM se describe el proceso jerárquico de esta metodología de la siguiente forma: En el nivel superior, el proceso de minería de datos se organiza en una serie de fases; cada fase consta de varias tareas genéricas de segundo nivel. Este segundo nivel pretende ser lo suficientemente amplio como para cubrir todas las situaciones posibles de minería de datos. Las tareas genéricas están diseñadas para ser lo más completas y estables posible. Al ser completas cubren tanto el proceso integral de minería de datos como todas sus aplicaciones posibles y al ser estables permiten que el modelo sea válido incluso para desarrollos imprevistos.

Correspondiente a las tareas especializadas, el tercer nivel permite describir cómo deben llevarse a cabo las acciones de las tareas genéricas en ciertas situaciones específicas. Por ejemplo, en el segundo nivel podría haber una tarea genérica llamada limpieza de datos. El tercer nivel describe cómo cambia esta tarea en diferentes situaciones, como la limpieza de valores numéricos frente a la limpieza de valores categóricos.

En el cuarto nivel se define la instancia del proceso, entendida como un registro de las acciones, decisiones y resultados de un proyecto real de minería de datos. Una instancia de proceso se organiza de acuerdo con las tareas definidas en los niveles superiores, pero representa lo que realmente sucedió en un compromiso particular, en lugar de lo que sucede de manera general.

La elección de CRISP-DM como la metodología con la que se ejecutará este proyecto aplicado se justifica por su capacidad para asegurar que cada fase esté alineada con los objetivos del negocio ya que permite abordar sistemáticamente desde la comprensión del problema y la preparación de los datos hasta la evaluación e implementación del modelo predictivo. Este enfoque permite integrar técnicas de análisis exploratorio, modelado estadístico y aprendizaje automático, facilitando la interpretación de resultados y su aplicación en decisiones estratégicas.

La metodología CRISP-DM se mantiene como un marco sólido y ampliamente aceptado para estructurar proyectos de minería de datos y analítica debido a su enfoque orientado al negocio, su flexibilidad y su aplicabilidad en múltiples dominios. Estudios revisados confirman que, incluso después de más de dos décadas desde su creación, sigue siendo el estándar de facto en proyectos tanto académicos como industriales (Schröer et al., 2021).

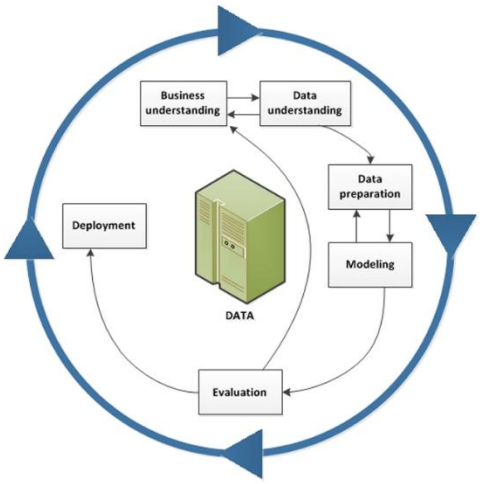
Todo el proceso de la metodología CRISP-DM es un proceso iterativo, el modelo se debe revisar periódicamente para verificar que aún se ajusta a las necesidades del negocio y que no es obsoleto, se necesita un monitoreo constante para determinar el mejor momento para pasar por el proceso nuevamente. Además, cuando se ejecuta el modelo CRISP-DM no se debe centrar en la etapa de modelado buscando pasar rápidamente a las otras etapas. Lo mejor es asegurarse que el proyecto tenga un enfoque claramente definido y que se tengan los datos correctos teniendo siempre clara la necesidad (Kelleher & Tierny, 2022).

**Etapas de la Metodología CRISP-DM**

La metodología CRISP-DM se compone de seis fases interrelacionadas que pueden iterarse según los hallazgos obtenidos en cada etapa (Chapman et al., 2000). Los datos se encuentran en el centro de las actividades de análisis de datos y las flechas entre los procesos indican la dirección del proceso.

**Figura 2**

*Fases del Modelo CRISP-DM*




*Nota*. Tomado de *Guía de CRISP-DM de IBM SPSS Modeler*, 2020, IBM.

### *Comprensión del Negocio*

Esta fase inicial busca definir los objetivos estratégicos del proyecto desde la perspectiva de la organización. Se traduce el problema de negocio en objetivos analíticos concretos, se evalúa la situación, se clarifican criterios de éxito y se diseña un plan preliminar para alcanzar los objetivos. Para esto es necesario explorar las expectativas de la organización respecto a la minería de datos implicando a la mayor cantidad de personas posibles. En el documento guía de esta metodología (Chapman et al., 2000) se explica que una comprensión deficiente del negocio puede llevar al desarrollo de soluciones técnicas que no tengan impacto real en la toma de decisiones.

En esta fase se realiza un análisis detallado del contexto, identificando las variables clave relacionadas con el desempeño empresarial en manufactura, en relación con el objetivo de apoyar la toma de decisiones. En esta etapa se busca entender los objetivos estratégicos de la organización que se está estudiando, precisar claramente el problema que se desea resolver y

definir el objetivo central del proyecto aplicado a la información de la Encuesta Anual Manufacturera del DANE. Según la guía de CRISP-DM, esta fase incluye la identificación de metas del negocio, criterios de éxito, restricciones y riesgos del proyecto. Una correcta comprensión del negocio es fundamental para garantizar que los resultados del análisis sean relevantes y útiles para la toma de decisiones (IBM, 2020). A continuación, se presentan las tareas genéricas de esta etapa (Chapman et al., 2000):

- Determinación de los objetivos comerciales: El primer objetivo es entender desde una perspectiva de negocio que quieren alcanzar los clientes y obtener la máxima información de los objetivos comerciales de la minería de datos. Para definir estos objetivos se deben desarrollar las siguientes etapas:

Compilación de la información de la empresa: Determinar la estructura de la organización, describir el área problemática y la solución actual.

Definición de los objetivos comerciales: Acordar un objetivo principal con los interesados en el proyecto, describiendo el problema que se desea resolver mediante la minería de datos, especificar todas las cuestiones comerciales de la forma más específica posible, determinar otros requisitos comerciales y especificar los beneficios esperados en términos comerciales.

Criterios de rendimiento comercial: Es importante definir la naturaleza del rendimiento comercial del proyecto de minería de datos.

- Evaluación de la situación: Esta tarea implica una investigación más detallada de todos los recursos, limitaciones, supuestos y otros factores que deben considerarse para determinar el objetivo del análisis de datos y el plan del proyecto. En esta tarea se debe profundizar en los detalles. ¿Qué tipos de datos están disponibles para el análisis? ¿Se dispone

del personal necesario para completar el proyecto? ¿Cuáles son los principales factores de riesgo? ¿Se dispone de planes de contingencia para cada factor de riesgo?

- Determinación de los objetivos de la minería de datos: Después de tener el objetivo comercial claro se debe traducir en un objetivo concreto de minería de datos.
- Producción de un plan de proyecto: Describir el plan previsto para alcanzar los objetivos de minería de datos y, por consiguiente, los objetivos de negocio. El plan debe especificar los pasos que se realizarán durante el resto del proyecto, incluida la selección inicial de herramientas y técnicas.

***Comprensión de los Datos***

Consiste en recolectar los datos disponibles y evaluar su calidad, estructura y relevancia para el problema planteado. En esta etapa se ejecuta un análisis exploratorio, identificación de valores faltantes, atípicos o inconsistentes.

En esta etapa se recopilarán los datos iniciales de la Encuesta Anual Manufacturera (EAM), incluyendo variables económicas, operativas y demográficas de las empresas (Departamento Administrativo Nacional de Estadística (DANE), 2025a). Se realiza un análisis exploratorio a los datos para familiarizarse con su estructura, contenido, calidad, variables, tipos de datos y calidad. Para de esta forma poder identificar las variables relevantes.

Esta etapa es crítica ya que una comprensión superficial de los datos puede resultar en modelos inadecuados o interpretaciones erróneas (Schröer et al., 2021). El documento guía señala que aquí se identifican problemas fundamentales como valores faltantes, valores atípicos (outliers), inconsistencias o patrones preliminares, lo que permite evaluar la idoneidad de los datos para el problema planteado (IBM, 2020). A continuación, se presentan las tareas genéricas de esta etapa, según la metodología original (Chapman et al., 2000):

- Recopilación de datos iniciales: En esta etapa debemos acceder a los datos provenientes de diversos orígenes.
- Descripción de los datos: En esta etapa se debe escribir un informe de descripción de datos y compartir los descubrimientos con los usuarios, esta descripción se debe hacer enfocándose en la cantidad y calidad de los datos.
- Exploración de datos: En esta fase se deben explorar los datos con tablas, gráficos y otras herramientas de visualización; la idea de este proceso es que este alineado a los objetivos de minería de datos, permita formular hipótesis y dar forma a las tareas de transformación de datos necesarias para análisis adicionales.
- Verificación de calidad de datos: Antes de proceder con el modelado es necesario verificar si existen errores de codificación, valores perdidos u otras incoherencias que afectan el análisis.

***Preparación de los Datos***

La preparación implica la limpieza, transformación, integración de fuentes y construcción del dataset final que será utilizado en el modelado. Se aplican métodos para gestionar valores faltantes, resolver inconsistencias, normalizar variables, depurar y transformar los datos y dividir el dataset en subconjuntos de entrenamiento y prueba con el fin de garantizar una base sólida y confiable para el análisis predictivo.

Esta etapa suele requerir más tiempo y recursos que el modelado mismo, debido a la complejidad del tratamiento de datos reales. Esta etapa es considerada una de las más demandantes del proceso. La guía de CRISP-DM enfatiza que una preparación adecuada de los datos es clave para el desempeño y la confiabilidad de los modelos analíticos (IBM, 2020). A continuación, se presentan las tareas genéricas de esta etapa (Chapman et al., 2000):

- Selección de datos: Decidir los datos que se utilizarán para el análisis de acuerdo con la relevancia para los objetivos de la minería de datos, la calidad y las limitaciones técnicas. Tener en cuenta que la selección de datos abarca tanto la selección de atributos (columnas) como la de registros (filas) en una tabla.
- Limpieza de datos: Elevar el nivel de calidad de los datos requerido para ejecutar las técnicas de análisis de datos e incluirlos en los análisis.
- Construcción de nuevos datos: Dependiendo de las necesidades del proyecto puede ser necesario construir nuevos datos a partir de derivación de atributos, generación de registros o valores transformados para atributos existentes.
- Integración de datos: Puede ser necesario integrar datos provenientes de varios orígenes de datos por medio del uso de un campo clave. Esta integración se puede ejecutar a través de la fusión o adición. También se combina información de múltiples tablas o registros para crear nuevos registros o valores.
- Formato de datos: Comprobar si algunas técnicas requieren aplicar un formato concreto o la clasificación de datos. Enfocándose en modificaciones sintácticas realizadas a los datos que no cambian su significado, pero que podrían ser requeridas por la herramienta de modelado.

***Modelado***

En esta fase se seleccionan y aplican técnicas de modelado cuantitativo basadas en aprendizaje supervisado para la estimación de la variable dependiente. En la etapa de modelado se debe seleccionar el algoritmo, el diseño de pruebas, la construcción del modelo y su respectiva evaluación. Dado que el problema se define como una regresión, se implementan técnicas robustas como árboles de decisión y Random Forest, las cuales permiten capturar relaciones no

lineales en los datos de la EAM. Durante esta etapa se ajustan los hiperparámetros y, de ser necesario, se retorna a la fase de preparación de datos para asegurar que los insumos cumplan con los requerimientos técnicos de cada algoritmo, garantizando así la mayor capacidad predictiva posible (Chapman et al., 2000). A continuación, se presentan las tareas genéricas de esta etapa:

- Selección de técnicas de modelado: Al seleccionar el tipo de modelado se debe tener en cuenta los tipos de datos disponibles, los objetivos de minería de datos y los requisitos de modelado.
- Generación de un diseño de comprobación: Definir como se comprobarán los resultados del modelo. Antes de construir un modelo, necesitamos generar un procedimiento o mecanismo para probar su calidad y validez.
- Generación de los modelos: Generar los modelos que se haya considerado. Utilizar la herramienta de modelado sobre el conjunto de datos preparado para crear uno o más modelos.
- Evaluación del modelo: Analizar los modelos detenidamente para determinar cuáles son los más precisos o eficaces para considerarlos finales. Interpretar los modelos según el conocimiento del dominio, los criterios de éxito de la minería de datos y el diseño de prueba deseado. Evaluar técnicamente el éxito de la aplicación de las técnicas de modelado y contactar con analistas de negocio y expertos en el tema para analizar los resultados de la minería de datos en el contexto empresarial.

***Evaluación***

Antes de implementar la solución, se evalúa si el modelo cumple con los criterios de éxito definidos inicialmente y si sus resultados son coherentes y útiles desde el punto de vista del

negocio. La guía oficial destaca que esta fase no se limita a métricas técnicas, sino que incluye una validación estratégica de los resultados obtenidos (IBM, 2020).

En la etapa de evaluación se da la interpretación de resultados en función de los objetivos del proyecto y se ejecuta un proceso de revisión. Antes de continuar, se deben evaluar los resultados obtenidos utilizando los criterios de rendimiento establecidos en el inicio del proyecto, lo cual es clave para asegurar que se puedan utilizar los resultados obtenidos.

Antes de implementar un modelo, se realiza una evaluación exhaustiva de su desempeño frente a los objetivos de negocio definidos inicialmente. No solo se analizan métricas técnicas, sino también la utilidad y aplicabilidad práctica de los resultados. Esta fase evita la implementación de modelos que, aunque técnicamente efectivos, no generen valor organizacional. A continuación, se presentan las tareas genéricas de esta etapa (Chapman et al., 2000):

- Evaluación de los resultados: Este paso evalúa el grado en que el modelo cumple con los objetivos de negocio y busca determinar si existe alguna razón empresarial que lo justifique. Otra opción es probar el modelo en entornos reales, si las limitaciones de tiempo y presupuesto lo permiten. Además, la evaluación también analiza otros resultados de minería de datos generados. Los resultados de minería de datos incluyen modelos que están necesariamente relacionados con los objetivos de negocio originales y todos los demás hallazgos que no están necesariamente relacionados con los objetivos de negocio originales, pero que también podrían revelar desafíos adicionales, información o pistas para futuras direcciones.
- Proceso de revisión: Proceso en el que se realiza una revisión de los aciertos y errores del proceso, se pretende aprender de la experiencia para que los proyectos de minería de datos sean más efectivos. Ahora es oportuno realizar una revisión más exhaustiva del trabajo de

minería de datos para determinar si se ha pasado por alto algún factor o tarea importante. Esta revisión también abarca cuestiones de control de calidad, por ejemplo: ¿Se construyó el modelo correctamente? ¿Se usaron solo los atributos permitidos y disponibles para futuros análisis?

- Determinación de los siguientes pasos: En esta fase se puede continuar con la fase de despliegue o volver y modificar o sustituir los modelos. Dependiendo de los resultados de la evaluación y la revisión del proceso, el equipo del proyecto decide cómo proceder. El equipo decide si finalizar el proyecto y pasar a la implementación, iniciar nuevas iteraciones o establecer nuevos proyectos de minería de datos. Esta tarea incluye el análisis de los recursos restantes y el presupuesto, lo cual puede influir en las decisiones.

***Despliegue***

La fase final corresponde a darle uso a los resultados del proyecto, ya sea mediante la implementación de modelos en producción, la generación de reportes de decisión o la automatización de procesos analíticos para implementar las mejoras en la organización. En esta etapa se implementan los resultados del proyecto, ya sea mediante la integración del modelo en sistemas productivos, la generación de informes o la formulación de recomendaciones. Según el documento CRISP-DM, el valor real del proyecto se materializa cuando los resultados analíticos se traducen en acciones concretas que apoyen la toma de decisiones organizacionales (Chapman et al., 2000).

En la etapa de despliegue se diseña la visualización de resultados, se produce un informe final y se revisa el proyecto. Finalmente, los resultados serán presentados para facilitar la toma de decisiones estratégicas y se proponen recomendaciones para futuros análisis y aplicaciones.

Aunque muchos estudios señalan que esta fase es la menos documentada en la literatura aplicada, su importancia radica en garantizar que los resultados se traduzcan en acciones

concretas que aporten valor a la organización (Schröer et al., 2021). A continuación, se presentan las tareas genéricas de esta etapa (Chapman et al., 2000):

- Planificación del despliegue: Planificar un despliegue completo y preciso de los resultados. Esta tarea toma los resultados de la evaluación y determina una estrategia de implementación. Si se ha identificado un procedimiento general para crear los modelos pertinentes, este se documenta para su posterior implementación.
- Planificación y control del mantenimiento: En un despliegue e integración completos de los resultados de modelado, el trabajo de minería de datos puede ser continuado. La monitorización y el mantenimiento son aspectos importantes si los resultados de la minería de datos se integran en el día a día del negocio y su entorno. La preparación cuidadosa de una estrategia de mantenimiento ayuda a evitar períodos innecesariamente largos de uso incorrecto de los resultados de la minería de datos. Para monitorizar la implementación de los resultados de la minería de datos, el proyecto necesita un plan detallado del proceso de monitorización que considere el tipo específico de implementación.
- Creación de un informe final: Al finalizar el proyecto, el equipo elabora un informe final que permite presentar los resultados a las personas interesadas. Dependiendo del plan de implementación, este informe puede ser solo un resumen del proyecto y sus experiencias o una presentación final y completa de los resultados de la minería de datos.
- Revisión final del proyecto: Formular las impresiones finales e incorporar los conocimientos adquiridos durante el proceso de minería de datos. Evaluar qué salió bien y qué salió mal, qué se hizo bien y qué necesita mejorarse.

## Análisis de la Operación Estadística: Encuesta Anual Manufacturera (EAM)

La Encuesta Anual Manufacturera (EAM) es la principal operación estadística de carácter estructural realizada por el DANE para el sector industrial en Colombia. Su propósito fundamental es generar información estratégica sobre la estructura, evolución y comportamiento del sector manufacturero, permitiendo el análisis de la producción, ventas, el empleo y la competitividad a nivel nacional y regional (Departamento Administrativo Nacional de Estadística (DANE), 2025a).

### Características de la Operación Estadística

- Tipo de operación: La operación estadística es de tipo censo.
- Periodicidad y alcance: Es una investigación de carácter anual. Tiene una cobertura nacional que incluye 30 departamentos, Bogotá D.C. y las principales áreas metropolitanas del país.
- Marco estadístico: Investigación estadística de carácter censal. El directorio base de fuentes a investigar se constituye a partir del directorio actualizado de la encuesta del año inmediatamente anterior. Se identifican los establecimientos objeto de estudio a partir de un marco estadístico que se genera con los directorios entregados por el Directorio Estadístico de la Dirección de Geoestadística.
- Cobertura geográfica: Esta encuesta tiene cobertura nacional para los establecimientos industriales manufactureros que cumplen con los requisitos de inclusión. Se produce información nacional, departamental y por áreas metropolitanas.
- Población objetivo: Establecimientos que se definen como industriales manufactureros según la CIIU Rev.4 A.C y que funcionan en el país. Estos establecimientos deben tener diez o más personas ocupadas o un valor de producción superior al estipulado

anualmente con el Índice de Precios del Productor (IPP) que para el año 2024 debe ser superior a $780,1 millones de pesos anuales.

- Unidad estadística: La unidad de observación y análisis es el establecimiento industrial.
- Desagregación geográfica: La EAM se ha hecho integrando información sobre los establecimientos industriales que funcionan en el país, de tal manera que los resultados obtenidos tengan cobertura nacional por áreas metropolitanas y departamentos.
- Desagregación temática: La información de la EAM se publica en la página web del Dane a nivel CIIU Rev. 4 A.C., a nivel de departamentos, áreas metropolitanas, organización jurídica, escala de personal y escala de producción. Para las materias primas y productos la información se encuentra bajo la clasificación central de productos CPC 2.1 A.C.
- Período de referencia: Periodicidad anual.
- Periodo de recolección: La recolección del operativo es anual y tiene una duración de alrededor de cinco meses (paralelamente con la captura, critica y validación) en el año siguiente al de referencia (t).
- Método de recolección: Formulario electrónico en un ambiente web por auto diligenciamiento y en caso de que se requiera plantilla de formulario en formato físico.

**Variables Principales**

Para entender el flujo de datos de esta encuesta, es vital comprender sus tres variables principales, las cuales definen el desempeño económico del sector:

- Producción bruta: El valor de la producción bruta es igual al "Valor de todos los productos, subproductos y trabajos manufacturados por el establecimiento en el año", más el "Valor de la energía eléctrica vendida", más el "Valor de las subvenciones causadas en el año",

más el "Valor de las existencias de los productos en proceso de fabricación al finalizar el año" menos el "Valor de los productos en proceso de fabricación al iniciar el año".

- Consumo intermedio: Representa el valor de los bienes y servicios (materias primas, empaques, energía, etc.) que se transforman o se consumen durante el proceso de producción de otros bienes y servicios.
- Valor agregado: Total de los ingresos recibidos por el uso de los factores productivos (tierra, capital, trabajo) participantes en el proceso de producción durante el período estudiado. El DANE calcula el valor agregado descontando del valor de la producción bruta el valor del consumo intermedio.

**Variables Específicas**

- Personal ocupado: Corresponde al personal que labora en la empresa o establecimiento (contratado de forma directa o a través de empresas especializadas) y a las personas propietarias, socias y familiares sin remuneración fija. También se diferencian por sexo y origen nacional.
- Remuneraciones: Sueldos, salarios y prestaciones sociales pagadas al personal ocupado.
- Inversión neta: Gastos destinados a la adquisición de activos fijos (maquinaria, equipo, edificaciones) menos las ventas de activos usados.
- Ventas y costos: Desglose de los ingresos por ventas de productos elaborados y los costos asociados a la mercancía vendida.
- Consumo de energía eléctrica (Kwh.): Corresponde a la cantidad de kilovatios (Kwh) de energía eléctrica que consume el establecimiento industrial durante el año.

La EAM es esencial para la toma de decisiones ya que permite identificar qué sectores están impulsando el crecimiento industrial. Permite obtener indicadores económicos del sector, entender la distribución, concentración o dispersión geográfica de la actividad de manufactura y generar estadísticas básicas para el cálculo de los agregados económicos de las cuentas nacionales. A partir de la información recolectada en esta operación estadística se puede entender las características y estado de la manufactura en el país para de esta forma tomar decisiones.

## Análisis Exploratorio y Selección de Variables Críticas

### Obtención de los Datos

La fase de obtención de datos para este proyecto aplicado se ejecutó con el aprovechamiento de fuentes oficiales, específicamente se usaron los datos obtenidos en la Encuesta Anual Manufacturera (EAM) producida por el DANE. El proceso se dividió en dos grandes etapas: la recolección original por parte de la entidad y el proceso de acopio y consolidación para la aplicación del modelo de aprendizaje.

### Proceso de Recolección y Acopio (Metodología DANE)

De acuerdo con el documento metodológico de la operación estadística Encuesta Anual Manufacturera (Departamento Administrativo Nacional de Estadística (DANE), 2025c) la recolección de los datos sigue un protocolo estandarizado para garantizar la calidad. A continuación, se presentan las etapas a través de las cuales se planea, organiza y desarrolla el operativo de campo para obtener la información; así como, las fases de tratamiento de los datos hasta terminar en la divulgación de resultados. Como actividades previas se realizan las siguientes acciones:

- Análisis de fuentes que requieren la aplicación de miniencuesta y la inclusión de nuevos establecimientos que se detectaron durante el año anterior.
- Revisión del directorio de la operación estadística con el área temática.
- Revisión técnica permanente del esquema de captura de la información.
- Actualización y adecuación de material de la encuesta.
- Adecuación y pruebas del aplicativo de captura de información.

La recolección de la información se realiza por auto diligenciamiento asistido a través de un formulario electrónico. El operativo se ejecuta en una sola fase de recolección de la

información a partir del día en que se realiza la notificación a las fuentes seleccionadas y hasta el cierre de la encuesta y entrega del informe final. De manera que al final de la operación se cuente con informes y se haga el cierre definitivo de la fase operativa para continuar con las labores de análisis y verificación de la información recolectada, con el fin de elaborar informes definitivos y realizar la difusión de estos. La duración para la etapa de recolección y el número de fuentes del directorio se define en el plan de recolección que se actualiza para cada período estadístico (Departamento Administrativo Nacional de Estadística (DANE), 2025c).

***Recolección***

La recolección se efectúa mediante formulario electrónico por auto diligenciamiento por parte de cada una de las fuentes de información incluidas en el directorio. La recolección se realiza a través de un sistema en línea alojado en servidores DANE, desarrollado en plataforma HTML y base de datos MQSL.

Al inicio de cada operativo se envía al establecimiento la notificación, en donde se le informa el esquema de recolección, se dan las instrucciones para realizar el ingreso al aplicativo, el proceso sobre cómo debe generar la contraseña para ingresar a través de la página web del DANE y se le explica la forma en que se debe reportar la información solicitada.

Las empresas cuentan con el acompañamiento de personal capacitado desde que reciben la notificación hasta que envían el formulario completamente diligenciado, para poder obtener la información con los parámetros de calidad y oportunidad establecidos. En los casos en que las empresas presenten inconvenientes en el diligenciamiento de la encuesta o no se evidencien registros de ingreso al formulario, se acude a realizar un contacto vía telefónica, correo electrónico, medios electrónicos o visita presencial.

Cuando la fuente ya ha diligenciado el formulario, el funcionario encargado revisa completamente la información y evalúa su consistencia. Para aprobar el formulario se deben solucionar todas las dudas relacionadas con inconsistencias o variaciones atípicas y; si es el caso, solicitar a la fuente realizar las correcciones o justificaciones correspondientes (Departamento Administrativo Nacional de Estadística (DANE), 2025c).

***Acopio***

El acopio de los datos se realiza una vez se cuenta con los formularios totalmente diligenciados, de esta manera se continua con el proceso de crítica de la información recolectada. Para ejecutar el proceso de crítica es importante tener en cuenta los siguientes aspectos:

- Crítica y codificación de la información recolectada: Este proceso consiste en realizar un análisis minucioso de la información reportada y una corrección o aclaración de los datos incongruentes o inconsistentes; de acuerdo con las normas y parámetros señalados en el instructivo de crítica de la información. Se pretende obtener la mayor precisión posible de los datos recolectados.
- El proceso de codificación de los productos y materias primas nuevas o aquellas que no fueron previamente diligenciadas se realiza con base en la Clasificación Central de Productos (CPC).
- Se realiza un análisis comparativo de los indicadores económicos, a nivel de cada formulario, con los del período inmediatamente anterior y con los resultados obtenidos en la Encuesta Mensual Manufacturera con Enfoque Territorial. El objetivo es analizar la información común a las dos operaciones estadísticas, para tener un control técnico de cada fuente en el tiempo.

- Consolidación y envío de información al DANE. Procedimiento que ejecuta el sistema y que consiste en consolidar y enviar vía electrónica al DANE la información de los formularios cuya información sea consistente con los parámetros metodológicos estipulados (Departamento Administrativo Nacional de Estadística (DANE), 2025c).

***Diseño de Procesamiento***

- Consolidación de archivos de datos: Los datos capturados por el sistema de auto diligenciamiento web por parte de las fuentes se encuentran consolidados directamente en la base de datos del DANE.
- Codificación: La EAM siguiendo las recomendaciones y lineamientos internacionales en la codificación industrial utiliza la CIIU Rev. 4 A.C., adaptada a las necesidades y requerimientos estadísticos a nivel nacional para codificar la información reportada por los establecimientos. Las materias primas y productos por reportar en la EAM se encuentran bajo la clasificación central de productos CPC 2.1 A.C. La codificación correspondiente a municipio y departamento se basa en la DIVIPOLA. Adicionalmente, se tienen otras variables con clasificaciones propias de la operación estadística como el tipo de encuesta, las novedades de campo, el estado de los establecimientos en el operativo, entre otras.
- Diccionario de datos: El diccionario de datos de la EAM contiene la descripción, características y validación de las variables utilizadas en la operación estadística.
- Revisión y validación: La revisión del archivo de datos inicia con las especificaciones de validación y consistencia que están activas al momento de diligenciar la información por parte de la fuente. Los formularios van cambiando de estado a medida que se van llevando a cabo cada una de las etapas desde la recolección hasta el proceso de aprobación en DANE central. Luego de que la información es diligenciada, se llevan a cabo los procesos de

crítica de los formularios diligenciados, donde inicialmente los analistas de las sedes del DANE a nivel nacional realizan un análisis de consistencia de la información diligenciada, se revisa la información histórica y la coherencia entre variables; teniendo en cuenta el contexto económico y las observaciones dadas por la fuente para las variaciones de las variables recolectadas. Con base en lo anterior, se realizan consultas adicionales a la fuente para verificar la consistencia de la información.

- Cuando los formularios revisados llegan a DANE central se realiza un proceso similar por parte del equipo de analistas con el fin de tener un doble punto de crítica; igualmente si es necesario de acuerdo con esta segunda revisión ampliar las observaciones por parte de la fuente el formulario es devuelto a la sede correspondiente para la indagación y complemento necesario. Por último, la información es analizada y revisada en el área temática de manufactura analizando las variaciones en las principales variables agregadas a los dominios de publicación y se realizan comités para verificar la consistencia por medio de análisis estadístico de variaciones y contribuciones.

- Diseño de instrumentos de edición (validación y consistencia) e imputación de datos: Se tiene un conjunto de normas y procedimientos para el tratamiento adecuado de la información, especificaciones de consistencia y validación para el aseguramiento de la calidad. Estas herramientas son utilizadas en la etapa de crítica, que se realiza para el formulario electrónico basándose en una ficha de análisis generada por el aplicativo de captura; de acuerdo con los principios metodológicos de la operación estadística descritos en el manual de crítica.

- Con respecto al subproceso de edición e imputación, a través de un formato establecido por temática, se solicita a las fuentes que omitieron el reporte de datos, la información que se deben imputar. Cuando se reciben estas solicitudes, se revisa la gestión

realizada e inicia el proceso de imputación de estas fuentes, para realizarlo, en la EAM se utilizan tres métodos de imputación. En el primer método, se revisa la Encuesta Mensual Manufacturera con Enfoque Territorial (EMMET) y se utiliza sus resultados para la imputación. El segundo método que se realiza, si no se cuenta con la información de EMMET, se toman los registros administrativos de Supersociedades que genera los datos de ingresos operacionales por empresa y PILA el cual contiene la información del personal ocupado a través de los pagos de seguridad social. El tercer método, se utiliza en caso de que la fuente no rinda información en la EMMET, se aplica la metodología del vecino más cercano y las variaciones de la clase a la que pertenece el establecimiento industrial.  Para imputar el módulo TIC y teniendo en cuenta que en su mayoría se trata de preguntas cualitativas, se replica la información del año inmediatamente anterior.

- Diseño para la generación de cuadros de resultados: De acuerdo con la información recolectada, para cada año se presenta la información de las variables principales clasificadas por diferentes conceptos, tales como: Actividad económica, escala de personal, escala de producción, departamento, área metropolitana y organización jurídica y por combinaciones de estos niveles.

- Las cifras definitivas se presentan en cuadros de salida, que contienen un análisis desagregado de las variables investigadas a los diferentes niveles, de acuerdo con especificaciones establecidas contenidas en el documento “Especificaciones de cuadros de salida” que hace parte de la documentación de la operación estadística. Se genera un cuadro denominado “Evolución variables principales según divisiones industriales CIIU Rev. 4 A.C”, para un total de 56 cuadros de salida. Se incluye un cuadro con la agrupación por tamaño de empresa y las principales variables de la EAM, teniendo en cuenta que la variable principal de

estudio de la EAM es la producción bruta y la desagregación de información está a nivel de establecimiento, por lo cual dicha desagregación se realiza sobre la variable producción bruta (Departamento Administrativo Nacional de Estadística (DANE), 2025c).

***Diseño del Análisis***

Comprende el análisis estadístico, el análisis de contexto y el comité de personas expertas. La coherencia de los resultados se establece mediante el análisis de los datos a nivel de microdato de cada uno de los capítulos de la encuesta y a nivel agregado de las variables de la EAM.

El cierre de la fase de recolección y acopio se realiza de acuerdo con el cronograma inicial establecido para la operación estadística, en el cual se estipulan entregas de bases preliminares, cuando se tiene un porcentaje de cobertura superior al 50%, con el fin de iniciar la revisión, validación y consistencia de la base de información, a través de chequeos de la información. Se inician los comités de consistencia de la información donde se revisan y se validan los datos de los establecimientos industriales, con el fin de detectar errores, datos atípicos, datos faltantes o errores de codificación y se realizan las imputaciones solicitadas. En el cronograma de la operación estadística se establece la fecha de entrega definitiva de la base consolidada, con la cual se inicia el procesamiento de la información, se genera una base definitiva depurada a partir de la cual se inicia la fase de análisis, donde se revisa la coherencia de los resultados, para esta fase el aplicativo cuenta con el módulo de análisis donde se puede revisar la coherencia de los resultados a través de consultas a nivel agregado y mediante la observación de la información a nivel de microdato de cada uno de los capítulos de la encuesta y a nivel agregado para las variables de la Encuesta Anual Manufacturera (EAM), que tiene como

resultado los cuadros de salida y productos de la operación estadística a publicar (Departamento Administrativo Nacional de Estadística (DANE), 2025c).

En el marco de la Encuesta Anual Manufacturera (EAM), cuando se identifican datos atípicos durante el proceso de análisis se solicita aclaración o revisión del dato en cuestión. Con el fin de validar la información, se debe confirmar la veracidad del dato y solicitar las evidencias correspondientes que sustenten su exactitud. Este proceso incluye el envío de documentación o reportes que permitan realizar los respectivos contrastes entre la información reportada y los parámetros normales de la encuesta, garantizando así la calidad y confiabilidad de los resultados.

- Métodos de análisis: Análisis de consistencia, análisis interno, análisis cruzado entre variables, análisis de contexto, análisis de comparabilidad, análisis estadístico, análisis univariado, análisis bivariado, análisis multivariado
- Anonimización de microdatos

Con el fin de asegurar la confidencialidad de los datos suministrados por las fuentes se entrega información en resúmenes numéricos que no hacen posible deducir de ellos información alguna de carácter individual que pudiera utilizarse para fines comerciales, de tributación fiscal, de investigación judicial o cualquier otro diferente del propiamente estadístico tal como lo exige la Ley 2335 de 2023, Ley de Estadísticas Oficiales de Colombia en su artículo 4. La anonimización se hace usando diferentes técnicas para agregar la información de las fuentes que se consideren identificables después de realizar un estudio de riesgo. Los usuarios de la información pueden acceder a las bases de datos a nivel de microdato sin variables de identificación. Con estas bases, quienes usan la información, generan las salidas que requieren y pueden disponer de los resultados, únicamente a nivel agregado, con el fin de cumplir con los

parámetros de reserva estadística (Departamento Administrativo Nacional de Estadística (DANE), 2025c).

- Verificación de la anonimización de microdatos

El proceso de anonimización de la EAM consiste en 3 macroprocesos: Aislamiento y armonización de la base de datos, identificación de escenarios de riesgo de la información de los establecimientos a partir de la cual se defina la aplicación de técnicas de anonimización y proceso de exportación de la base en formato Excel para su revisión y proteger la privacidad de las fuentes preservando el aprovechamiento de los datos (Departamento Administrativo Nacional de Estadística (DANE), 2025c).

- Comités de expertos

En la EAM se realiza un análisis de los resultados al interior de la misma operación estadística sobre toda la información disponible, para estudiar su evolución y dinámica frente a la economía nacional. Para la producción y análisis de los resultados se cuenta con el comité interno con la participación de representantes del equipo de trabajo involucrado en esta operación estadística. Como resultado de estas reuniones surgen ejercicios que permiten explicar a profundidad los resultados. Una vez avaladas las cifras en la institución se socializan los resultados obtenidos en el año de estudio en un comité externo de expertos al que pertenecen representantes del Banco de la República, el Departamento Nacional de Planeación (DNP), el Ministerio de Hacienda, Ministerio de Industria y Comercio, Ministerio de Minas y Energía y la Asociación Nacional de Empresarios de Colombia (ANDI), entre otros; quienes están al tanto de los resultados de la operación, dando aplicación a los principios del código nacional de buenas prácticas para las estadísticas oficiales en el país, garantizando siempre la reserva estadística (Departamento Administrativo Nacional de Estadística (DANE), 2025c).

- Evaluación del análisis

Revisados y aprobados los resultados de la información que se va a difundir, se evalúan los métodos de análisis empleados y las actividades desarrolladas. Se plantean acciones de mejora para el siguiente año de estudio de la EAM.

**Descripción del Dataset**

El universo de estudio de la operación estadística Encuesta Anual Manufacturera está constituido por establecimientos industriales con 10 o más personas ocupadas o que superen el valor de producción estipulado anualmente (Departamento Administrativo Nacional de Estadística (DANE), 2025b). En esta operación estadística se recolecta información de caracterización inicial del establecimiento, personal contratado (Desagregado por tipo de contratación, nivel educativo y sexo), salarios y prestaciones sociales del personal contratado, costos y gastos (Administrativos y operacionales), energéticos (Consumidos y vendidos), producción (Productos manufacturados, cantidades producidas, cantidades vendidas, valor de producción y valor de venta), materias primas (Materias primas consumidas, cantidades consumidas, cantidades compradas, valor de consumo y valor de compra) y otros ingresos adicionales. A partir de este conjunto de variables, se realizó una selección estratégica de las siguientes variables siguiendo criterios de relevancia económica y capacidad predictiva:

- Variable objetivo: Valor agregado (Indicador de riqueza generada).
- Variables predictoras:
  - Producción Bruta (Escala operativa).
  - Consumo Intermedio (Eficiencia de costos).
  - Personal Permanente (Factor trabajo).
  - Activos Fijos (Capacidad instalada).

- Actividad económica (Patrón de comportamiento).

Esta selección permite que el modelo cumpla con el requisito de identificar los patrones iniciales y la viabilidad técnica, porque estas variables cuentan con una integridad superior al 99% en el reporte oficial del DANE. Ya que al realizar un resumen de la calidad de los datos se verificó que en las variables a analizar no se encontraron valores nulos.

**Tabla 1**

*Diagnóstico de Calidad de Datos*

| Variable | Tipo de Dato | Registros Nulos | Integridad | Total Registros |
|---|---|---|---|---|
| Valor_Agregado | int64 | 0 | 100.00% | 13297 |
| Prod_Bruta | int64 | 0 | 100.00% | 13297 |
| Consumo_Interm | int64 | 0 | 100.00% | 13297 |
| Sueldos_Pers_Perm | int64 | 0 | 100.00% | 13297 |
| Inv_Activos | int64 | 0 | 100.00% | 13297 |
| CIIU | int64 | 0 | 100.00% | 13297 |
| Depto | int64 | 0 | 100.00% | 13297 |

*Nota*. Diagnóstico de la calidad de los datos.

## Caracterización de las Variables

Para este proyecto aplicado se han seleccionado variables que representan los pilares fundamentales de la actividad industrial:

- Valor agregado (Variable objetivo): Es el total de los ingresos recibidos por el uso de los factores productivos (Tierra, capital, trabajo) participantes en el proceso de producción durante el período estudiado. Es el excedente económico generado por la transformación de insumos en productos y representa el valor real que la empresa aporta a la economía (Departamento Administrativo Nacional de Estadística (DANE), 2025c). El DANE calcula el valor agregado como la diferencia entre la Producción Bruta y el Consumo Intermedio.

- Producción bruta: El valor de la producción bruta es igual al "Valor de todos los productos, subproductos y trabajos manufacturados por el establecimiento en el año", más el "Valor de la energía eléctrica vendida", más el "Valor de las subvenciones causadas en el año", más el "Valor de las existencias de los productos en proceso de fabricación al finalizar el año", menos el "Valor de los productos en proceso de fabricación al iniciar el año". Producción bruta=Producción industrial+ (Inventario final de productos en proceso - Inventario inicial de productos en proceso) + Valor de energía eléctrica vendida+ Valor de subvenciones. Es el valor total de la producción y servicios generados por el establecimiento e indica el tamaño operativo del establecimiento (Departamento Administrativo Nacional de Estadística (DANE), 2025c).
- Consumo intermedio: Representa el valor de los bienes y servicios no durables utilizados como insumos en el proceso de producción para elaborar productos y servicios manufactureros (Departamento Administrativo Nacional de Estadística (DANE), 2025c).
- Número de empleados permanentes: Cuantifica la contratación de fuerza laboral permanente, indicando la intensidad del factor trabajo usada por el establecimiento industrial en el desarrollo de las actividades de manufactura (Departamento Administrativo Nacional de Estadística (DANE), 2025c).
- Valor de activos fijos: Refleja la inversión en maquinaria y equipo industrial y edificaciones (capacidad instalada) realizada por el establecimiento (Departamento Administrativo Nacional de Estadística (DANE), 2025c).

A partir de la selección de las anteriores variables se busca desarrollar un modelo que permita modelar y caracterizar la variable valor agregado de las empresas manufactureras, identificando la eficiencia en la transformación de recursos y la capacidad de generación de riqueza del sector manufacturero. Los objetivos de las variables seleccionadas son los siguientes:

- Las variables de producción bruta y consumo intermedio actúan como indicadores directos de la escala operativa.
- Las variables de personal permanente y activos fijos sirven para medir el impacto del capital humano y la infraestructura en los resultados económicos del establecimiento.

**Estadísticas Descriptivas**

A continuación, se presenta la tabla de caracterización estadística de los establecimientos industriales estudiados en los años 2023 y 2024 enfocada en las variables clave seleccionadas.

**Tabla 2**

*Estadísticas Descriptivas*

| Variable | Conteo | Promedio | Des. Estándar | Mínimo | P25 | Mediana | P75 | Máximo | CV (%) |
|---|---|---|---|---|---|---|---|---|---|
| Producción Bruta | 13297.00 | 66318023.05 | 292344587.84 | 2604.00 | 2032047.00 | 7219609.00 | 35551426.00 | 9131524236.00 | 440.82 |
| Valor Consumo | 13297.00 | 43973751.93 | 258028998.64 | 1486.00 | 929643.00 | 3486211.00 | 18992532.00 | 8325687836.00 | 586.78 |
| Sueldos Personal | 13297.00 | 2522005.80 | 5581101.61 | 0.00 | 213624.00 | 647081.00 | 2276023.00 | 61493268.00 | 221.30 |
| Activos Fijos | 13297.00 | 94086.91 | 1404747.59 | 0.00 | 0.00 | 0.00 | 0.00 | 79382305.00 | 1493.03 |
| Valor Agregado | 13297.00 | 22344271.12 | 61933127.50 | 5.00 | 890860.00 | 3235353.00 | 14467277.00 | 1039594116.00 | 277.18 |

*Nota*. Estadísticas descriptivas de la base de datos.

***Análisis de Heterogeneidad***

El dato más significativo es el *CV (%)* (Coeficiente de Variación). Se observa que todas las variables tienen un coeficiente de variación superior al 220% y en el caso de las Inversiones en Activos Fijos, llega a un 1493%. Esto confirma que la industria manufacturera en Colombia es heterogénea. Se tienen establecimientos muy pequeños y establecimientos con grandes plantas

de producción. Esta volatilidad estadística justifica técnicamente la aplicación de escalamiento y normalización de datos, garantizando que las magnitudes extremas no sesguen el aprendizaje del modelo ni invaliden la precisión de las predicciones.

***Análisis de Medidas de Posición***

Respecto a la variable valor agregado el promedio es de 22.3 millones, pero la mediana (50%) es de apenas 3.2 millones. Esto indica que los datos están sesgados a la derecha. Es decir, la gran mayoría de las empresas generan un valor cercano a los 3 millones, pero hay unas pocas empresas extremadamente grandes (el máximo es de 1.039 millones) que empujan el promedio hacia arriba. Esto hace que la mediana sea una medida mucho más realista de la empresa característica.

***Representatividad de la Muestra***

El conteo de 13.297 registros es un número robusto. Al tener más de 13 mil observaciones integradas de los periodos 2023-2024, el modelo tiene suficientes datos para aprender patrones estadísticos confiables sin caer en el sobreajuste (overfitting).

***Diagnóstico por Variable***

- Valor de la producción bruta: Es la variable con los valores más altos (máximo de 9.131 millones). Su coeficiente de variación (440.82%) es alto, lo que confirma una gran desigualdad; es decir, existen empresas que producen miles de millones frente a empresas con bajos niveles de producción, situación que justifica el uso de modelos no lineales como Random Forest.
- Total de inversiones en activos fijos: El P25, la mediana (50%) y el P75 son 0.0. Esto indica que la gran mayoría de las empresas no realizaron inversiones en terrenos o activos fijos (Maquinaria y equipo industrial) durante los periodos reportados. Solo un grupo muy

pequeño (el 25% superior) concentra toda la inversión. El coeficiente de variación de 1493% muestra que solo un grupo muy pequeño de empresas grandes realizo algún tipo de inversión en infraestructura y equipo industrial.

- Total valor consumo intermedio: Son los gastos operativos necesarios para producir. Al tener un promedio de $4.3M, nos dice que la eficiencia operativa varía drásticamente entre sectores, impactando directamente en qué tanto valor se queda en la empresa.
- Total sueldos y salarios del personal permanente: Es la inversión directa en el talento humano estable de la industria. Es la variable más estable de todas ya que el coeficiente de variación de 221% es el más bajo del grupo de variables estudiadas. Esto indica que, independientemente de si a la empresa le va muy bien o muy mal en ventas, los compromisos salariales se mantienen constantes.
- Valor agregado: Es la riqueza neta creada por el establecimiento, lo que queda después de pagar los factores que se usan directamente en el proceso productivo. Esta variable presenta una marcada asimetría positiva, evidenciada por la brecha entre el promedio de $22.3M y la mediana de $3.2M. Este comportamiento revela que la estructura de la industria manufacturera analizada está compuesta mayoritariamente por establecimientos de pequeña escala que generan niveles moderados de valor agregado, mientras que una minoría de grandes establecimientos concentra la mayor parte del valor agregado creado, elevando significativamente el promedio aritmético.

La caracterización estadística revela una industria manufacturera diferenciada por la alta dispersión de sus magnitudes económicas, con un coeficiente de variación del 277% en el valor agregado. La marcada diferencia entre los promedios y las medianas de producción bruta y

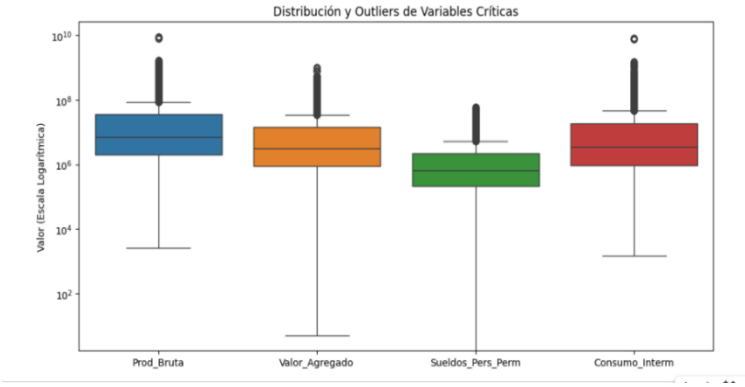
sueldos y salarios del personal permanente evidencia una estructura donde un pequeño grupo de establecimientos concentra la mayor parte de la capacidad productiva.

Estos hallazgos validan la decisión técnica de emplear algoritmos basados en árboles de decisión (Random Forest), dada su robustez frente a distribuciones con sesgo y valores extremos, permitiendo una predicción del valor agregado que reconoce la realidad tanto de las medianas empresas como de los grandes establecimientos. Además, permite ver que se está trabajando con un sector industrial sumamente diverso y desigual, lo cual justifica técnicamente por qué un modelo simple no bastaría y por qué el Random Forest se debe implementar.

A continuación, se presenta un gráfico de Boxplot que permite analizar las variables seleccionadas:

**Figura 3**

*Gráfico Boxplot de Variables Criticas*




*Nota*. Caracterización de las variables seleccionadas.

- Jerarquía de magnitudes (Producción bruta vs. Consumo intermedio)

Al observar la caja azul (Producción bruta) frente a la caja roja (Consumo intermedio), se puede verificar que la producción bruta es la más alta, pero la proximidad de la caja roja (Consumo intermedio) confirma que la industria manufacturera colombiana es intensiva en

insumos. Una gran parte del valor de lo que se produce se usa directamente en la compra de materias primas necesarias en los procesos productivos.

- Valor agregado (Caja naranja)

El valor agregado es la diferencia entre la producción bruta (Caja azul) y el consumo intermedio (Caja roja). En el gráfico, la caja naranja se sitúa en un nivel intermedio. Representa la riqueza real que la industria manufacturera añade a los materiales de acuerdo con sus actividades de producción. El hecho de que la caja naranja sea más alta que la caja verde (Sueldos del personal permanente) indica que la industria genera suficiente valor para cubrir los costos y gastos de personal y generar un margen operativo considerable para otros costos y utilidades.

- Sueldos y salarios (Caja verde)

La caja verde es la más baja y la que tiene una dispersión ligeramente menor en su cuerpo principal. Los salarios son el costo más controlado y estable. Sin embargo, los outliers (puntos negros) en esta columna son vitales ya que muestran que hay un pequeño grupo de empresas con una alta inversión en sueldos y salarios, que son justamente las que disparan la producción hacia arriba.

- Outliers (Puntos negros)

En las cuatro columnas, los outliers se extienden mucho más allá de las cajas. Esto confirma que el sector manufacturero no es un conjunto uniforme. Hay un conjunto de empresas (los puntos superiores) cuyas cifras de producción y consumo son mayores que las de una empresa promedio.

El análisis a partir del grafico de Boxplot permite visualizar la estructura de costos de la industria: la producción bruta (Caja azul) está fuertemente impulsada por el consumo intermedio

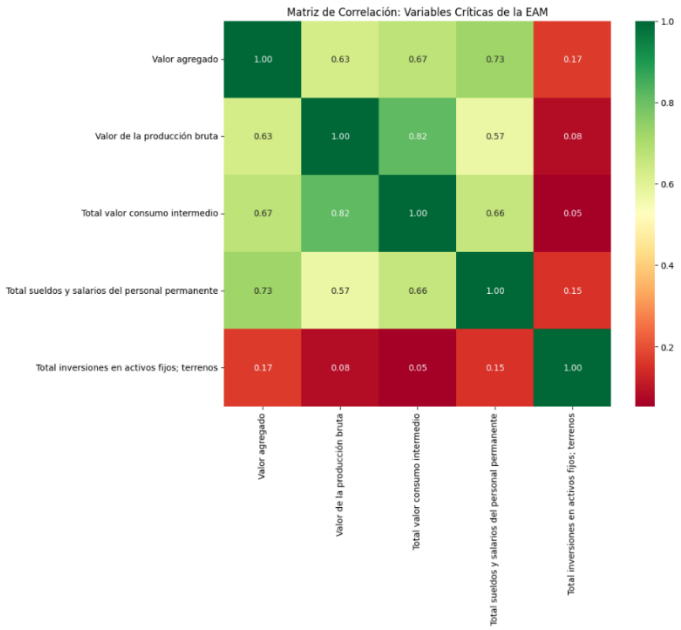
(Caja roja), mientras que los sueldos (Caja verde) representan el componente más estable, pero de menor magnitud relativa. La brecha entre la producción y el consumo se consolida en el valor agregado (Caja naranja). La presencia masiva de valores atípicos en todas las dimensiones ratifica la asimetría del sector y justifica el uso de algoritmos capaces de modelar tanto el comportamiento de la industria media como el de los grandes establecimientos industriales atípicos.

**Análisis de la Matriz de Correlación**

La matriz de correlación de Pearson permite cuantificar la fuerza y la dirección de la relación lineal entre las variables numéricas del proyecto. Los valores oscilan entre -1 (Correlación negativa perfecta) y 1 (Correlación positiva perfecta).

**Figura 4**

*Matriz de Correlación (Variables Críticas)*




*Nota*. Correlación entre las variables seleccionadas.

- Relación crítica (Valor agregado vs. Producción Bruta): Se observa una correlación entre el valor agregado y la producción bruta de 0.63. Esto indica que el valor agregado es extremadamente sensible al volumen de producción. Esta relación sugiere que la industria no solo genera riqueza por volumen de venta, sino por la transformación eficiente que hace el personal sobre los insumos. Desde el punto de vista predictivo, la producción bruta será el motor principal del algoritmo.
- Eficiencia de insumos (Valor agregado vs. Consumo Intermedio): La correlación entre el valor agregado y el consumo intermedio es de 0.67. Esto sugiere que no todo el aumento en el gasto de insumos y elementos usados en el proceso de producción se traduce linealmente en un aumento del valor agregado, lo que deja espacio para que el modelo identifique ineficiencias operativas.
- Sueldos y salarios del personal permanente (Valor agregado vs. Sueldos y Salarios Personal Permanente): Se presenta una correlación de 0.73 entre el valor agregado y los sueldos y salarios del personal permanente. Esto confirma que el talento humano es un predictor significativo. Existe una relación directa y potente entre el pago de salarios y la generación de valor agregado. Esto indica que la industria colombiana analizada es altamente dependiente del talento humano. Es interesante notar que los sueldos y salarios del personal tienen una correlación más baja con la producción bruta (0.57) que, con el valor agregado, lo que implica que el personal está más directamente ligado a la rentabilidad final que al proceso de fabricación.
- Capacidad instalada (Valor agregado vs. Inversión en activos fijos): El valor agregado y la inversión en activos fijos tienen la correlación más baja del grupo (0.17). Esto no significa que no sea importante, sino que la relación entre la inversión en maquinaria y equipo industrial y el valor agregado no es tan directa o lineal como la producción. Los activos fijos

suelen generar valor a largo plazo o de forma escalonada. Esto nos dice que, en el corto plazo comprar terrenos o activos fijos no garantiza un aumento inmediato en el valor agregado. Las inversiones en infraestructura suelen dar resultados a largo plazo, por lo que, en este análisis de solo dos años, su impacto parece ruido para el modelo.

La matriz revela que existe multicolinealidad entre las variables ya que las variables predictoras también están muy relacionadas entre sí, por ejemplo, producción bruta y consumo intermedio con 0.82. Esta situación permite ver que el modelo será muy preciso para predecir, pero debemos tener cuidado al decir qué variable es más importante que otra de forma aislada, ya que todas actúan como un sistema integrado dentro del establecimiento industrial.

El diagnóstico mediante la matriz de correlación de Pearson valida la selección de variables, evidenciando una asociación positiva y fuerte entre los factores de producción y el valor agregado. La alta correlación de la producción bruta (0.63) y el consumo intermedio (0.67) con el valor agregado asegura que el modelo contará con señales claras para el aprendizaje supervisado, confirmando la viabilidad técnica del análisis predictivo propuesto.

**Diagnóstico de Viabilidad**

Tras ejecutar el diagnóstico, se observa que la integridad de las variables clave es superior al 99%. Al no existir una falta crítica de datos en las variables principales y al existir correlación entre estas variables, se concluye que la base de datos es técnicamente viable para el entrenamiento de modelos de Machine Learning. Tras el análisis exploratorio, se confirma la viabilidad del proyecto basándose en tres pilares:

- Existencia de correlación: La variable objetivo (Valor agregado) presenta una relación lineal fuerte con los sueldos y salarios del personal permanente (0.73) y la producción bruta (0.63), lo que garantiza que los algoritmos encontrarán patrones predictivos claros.

- Complejidad del sistema: El alto coeficiente de variación (CV=440%) en la producción bruta indica que una regresión lineal será insuficiente, justificando el uso de algoritmos de ensamble como Random Forest.
- Volumen de datos: Con más de 13,000 registros consolidados, la muestra es estadísticamente significativa para entrenar modelos robustos sin caer en el sobreajuste (overfitting).

## Procesamiento de los Datos

### Consolidación de la Base de Datos

Para los fines de este proyecto aplicado de aprendizaje supervisado, se realizó un proceso de consolidación de los datos recolectados en la operación estadística Encuesta Anual Manufacturera para los años 2023 y 2024:

- Extracción: Se descargaron los archivos maestros en formato Excel correspondientes a los microdatos anonimizados de ambos años.
- Preparación técnica: Se utilizó un script de Python para localizar la estructura de datos a partir de la fila de encabezados (identificada técnicamente en la fila 4 de los archivos fuente), asegurando la lectura correcta de los nombres de las variables.
- Unificación temporal: Se ejecutó una consolidación de ambas bases de datos para construir un dataset robusto que permitiera capturar la variabilidad económica, facilitando así la identificación de patrones de crecimiento en el valor agregado. Esta base de datos consolidada tiene 13297 registros.

**Tabla 3**

*Datos Consolidados*

| | ID_Empresa | ID_Establecimiento | CIIU | Prod_Bruta | Consumo_Interm | Sueldos_Pers_Perm | Valor_Agregado |
|---|---|---|---|---|---|---|---|
| 0 | 141160 | 140162 | 1081 | 1316771 | 917919 | 236218 | 398852 |
| 1 | 141160 | 145341 | 1081 | 7598434 | 4281772 | 1220125 | 3316662 |
| 2 | 141162 | 140163 | 2212 | 2549153 | 2423276 | 240844 | 125877 |
| 3 | 141163 | 141099 | 2011 | 41736224 | 13840803 | 2244356 | 27895421 |
| 4 | 141164 | 140166 | 1410 | 1309557 | 1104870 | 114000 | 204687 |

*Nota*. Primeros datos de la base de datos consolidada.

Después de realizar el proceso de consolidación de la información se ejecutó un análisis exploratorio descriptivo, el cual permite establecer la línea base del proyecto, permitiendo

identificar una alta volatilidad estructural en el sector. Estos resultados validan la necesidad de un preprocesamiento de datos riguroso y sustentan la elección de modelos de aprendizaje supervisado capaces de gestionar distribuciones con sesgo positivo y alta presencia de valores atípicos. También se realizó una caracterización de estos datos buscando determinar la presencia de valores nulos que afecten el proceso de implementación del modelo.

Las variables *nordem* y *nordest* del establecimiento se descartarán en la fase de modelado ya que son un identificador único que carecen de significancia estadística para la predicción de magnitudes económicas. Su exclusión previene el sobreajuste (*overfitting*) de los algoritmos y asegura que las predicciones se fundamenten en factores de producción reales (Capital y trabajo) y no en códigos de identificación, cumpliendo además con los principios de anonimización de microdatos. La variable de actividad económica (CIIU) se descartará ya que muestra la actividad industrial que ejecuta el establecimiento y su inclusión no es indispensable para capturar la heterogeneidad estructural de la industria manufacturera a través de la variable valor agregado.

**Preparación de los Datos**

- Mapeo de variables: Se implementó una función de limpieza semántica para traducir los códigos técnicos a nombres descriptivos que permitan entender la variable que se está estudiando. Esto asegura que los resultados del modelo sean interpretables por el usuario de la información.

- Tratamiento del tipo de datos (String a numérico): Las variables monetarias se aseguraron como tipo float64 para permitir cálculos de alta precisión. Se identificaron variables como *dpto* y *ciiu4* para ser tratadas como categóricas (strings) en fases posteriores, evitando que el modelo les asigne un valor numérico erróneo.

- Eliminación de columnas: Se descartaron los identificadores anónimos (*nordemp*, *nordest*) para evitar el overfitting, garantizando que el modelo aprenda patrones industriales y no memorice registros específicos.
- Integración temporal: Se consolidaron los datos recolectados para el año 2023 y el año 2024. Esta decisión fortalece el modelo al proporcionarle una muestra más amplia y variada, capturando la dinámica económica de dos períodos distintos.
- Validar calidad: Se ejecutó un conteo de valores nulos y outliers para entender las características del dataset.

**Limpieza de Outliers y Valores Nulos**

Dado que el sector manufacturero colombiano es heterogéneo (coexisten microempresas y grandes industrias), para asegurar la robustez y la capacidad de generalización del modelo de regresión, se realizó un proceso de depuración estadística sobre la variable objetivo (Valor agregado) por medio de la limpieza de outliers. Para el desarrollo de este proceso se aplicó el método de Rango Intercuartílico (IQR), en el cual se consideraron outliers aquellos valores que se encontraban por debajo del percentil 25 menos 1.5 veces el rango intercuartílico o por encima del percentil 75 más 1.5 veces el rango intercuartílico.

Como resultado de este procedimiento, se identificaron y eliminaron 1,846 registros que presentaban valores extremos en la variable valor agregado, lo que representa una reducción del 13.88% de la muestra inicial (pasando de 13,297 a 11,451 observaciones finales). Este ajuste permitió eliminar el ruido estadístico provocado por casos excepcionales, logrando que los algoritmos a implementar se enfoquen en los patrones de productividad representativos de la gran mayoría de los establecimientos industriales, mejorando así la precisión de la estimación y reduciendo el sesgo en el entrenamiento.

Debido a que en la EAM hay empresas pequeñas y empresas muy grandes esta técnica permite eliminar registros con valores atípicos extremos que podrían sesgar las predicciones, asegurando que el modelo aprenda el comportamiento industrial normal y no se distraiga con casos excepcionales que generan ruido. Luego se seleccionaron las variables separando la variable objetivo de las variables predictoras, estas variables se transforman identificando cuales son numéricas y cuales son tipo texto.

**Trazabilidad del Procesamiento de los Datos**

La siguiente tabla resume el flujo de transformación de los datos, desde la fuente original hasta el conjunto de entrenamiento final:

**Tabla 4**

*Trazabilidad del Proceso de Limpieza*

| Etapa del Proceso | Registros (N) | Registros Eliminados | Justificación Técnica |
|---|---|---|---|
| Base de datos año 2023 | 6,714 | 0 | Total de establecimientos que reportaron información en el año 2023 |
| Base de datos año 2024 | 6,583 | 0 | Total de establecimientos que reportaron información en el año 2024 |
| Base inicial (EAM) | 13,297 | 0 | Punto de partida con la totalidad de establecimientos reportados por el DANE. |
| Limpieza de Outliers | 13,297 | 1,846 | Aplicación del método IQR (1.5x) sobre el valor agregado para mitigar el sesgo provocado por la alta heterogeneidad y mejorar la capacidad de generalización del modelo. |
| Muestra final para modelado | 11,451 | 1,846 | Tamaño de muestra definitivo que garantiza un soporte estadístico robusto para el entrenamiento de algoritmos de ensamble. |

*Nota*. Resumen del proceso de limpieza ejecutado.

A continuación, se presentan las decisiones tomadas durante la ejecución del proceso de consolidación de los datos:

- Consolidación de la base de datos: Se trabajó sobre los datos de la Encuesta Anual Manufacturera del año 2023 y del año 2024. Luego se realizó la consolidación de estos dos archivos en un único dataset para generar la base inicial de 13,297 establecimientos que representa el universo completo de la industria en los dos periodos analizados.
- Criterio de exclusión (Outliers): La eliminación del 13.88% de registros de la muestra no responde a una limpieza aleatoria, sino a la necesidad técnica de los modelos de regresión de trabajar con distribuciones menos sesgadas. En el sector manufacturero, los valores extremos (outliers) suele corresponder a grandes establecimientos cuyas estructuras de costos no son comparables con el promedio nacional. Mantenerlos elevaría el Error Medio Absoluto (MAE), restando utilidad al modelo para la mayoría de los establecimientos.
- Representatividad de la muestra final: A pesar de la limpieza, el número final de 11,451 registros excede ampliamente los requisitos mínimos para el entrenamiento de un modelo de Random Forest. Esto permite que el modelo aprenda patrones de eficiencia reales sin verse afectado por el ruido de los casos excepcionales, cumpliendo así con el objetivo de generar una herramienta de estimación prospectiva confiable.

**Imputación de Datos Faltantes**

La imputación de datos faltantes es un proceso que consiste en reemplazar los valores faltantes por valores estimados. Los datos faltantes pueden deberse a errores en la recolección, problemas en la transmisión o en el almacenamiento de los datos.

Se realizó un análisis para la identificación de datos faltantes, a partir del cual se concluyó que en la base de datos consolidada no existen datos faltantes en las variables clave por lo cual no fue necesario realizar actividades de imputación.

**Transformación y Escalamiento de Variables**

Para que los algoritmos de aprendizaje automático procesen la información de manera equitativa, se ejecutó la transformación de las variables mediante la técnica de estandarización (Standard Scaling). Este método se caracteriza porque aplicó un escalador estándar a las variables monetarias y de personal a una escala común, con el cual se evita que variables con números muy grandes (como producción bruta) sesguen el modelo o resten importancia a variables con rangos numéricos menores (como el salario de personal personal). De esta forma, se asegura que cada predictor aporte un peso relativo justo en la estimación del valor agregado (Massaron & Boschetti, 2016).

**Descripción de la Base de Datos Final**

Esta sección detalla la base de datos final resultante del proceso de preprocesamiento y selección de variables, la cual fue utilizada para realizar el entrenamiento y validación de los modelos predictivos seleccionados.

El conjunto de datos consolidado con el cual se realizará el modelado está compuesto por 11,451 registros, obtenidos después de realizar la depuración de la base consolidada de los archivos históricos de la base de datos de la EAM del año 2023 y de la EAM del año 2024. Los archivos fuente se publicaron en formato xslx y fueron unificados en un único archivo para optimizar la lectura, compresión y acceso columnar.

Cada fila del conjunto de datos consolidado representa un establecimiento que reportó la información solicitada en los dos años de estudio. Cada una de las columnas representa una variable de estudio de la información recolectada.

Para enfrentar los desafíos derivados de la magnitud del conjunto de datos y su variabilidad temporal, se implementó una estrategia de procesamiento que permitió explorar y

documentar las variables sin comprometer la estabilidad del sistema que procesa los datos. En el marco de esta estrategia se utilizó un diccionario de datos en el que se puede verificar las variables y su respectivo nombre.

La limpieza y estandarización de los microdatos de los años 2023 y 2024 han permitido construir una base de datos sólida y libre de ruido estadístico, cumpliendo con los requisitos técnicos necesarios para proceder a la fase de modelado predictivo. Este diagnóstico confirma la necesidad de aplicar, en fases posteriores, estrategias de recodificación, filtrado por periodos y estandarización de variables, con el fin de garantizar la calidad y la comparabilidad histórica de la base de datos. Finalmente se puede observar que en las columnas no existen datos faltantes o nulos.

El conjunto de datos final se compone de las siguientes columnas, las cuales fueron seleccionadas por su relevancia y su aporte al modelado predictivo. A continuación, se presentan las variables seleccionadas:

- Prod_Bruta: Producción bruta generada por el establecimiento.
- Consumo_Interm: Consumo intermedio generado por el establecimiento.
- Sueldos_Pers_Perm: Sueldos y salarios del personal permanente contratado.
- Inv_Terrenos: Inversión en activos fijos y maquinaria y equipo industrial.
- Valor_Agregado: Valor agregado generado por el establecimiento.

# Entrenamiento del Modelo

Este capítulo describe el proceso seguido para construir y entrenar los modelos de aprendizaje automático necesarios para entender las relaciones de los microdatos de la Encuesta Anual Manufacturera. A partir de los datos preparados y divididos, se procedió a seleccionar algoritmos de aprendizaje automático con distintos niveles de complejidad y enfoques de modelado. El objetivo es comparar el desempeño de los algoritmos seleccionados bajo condiciones homogéneas y seleccionar el modelo más adecuado para el problema planteado en este proyecto aplicado. Además, se implementó un esquema de optimización de hiperparámetros para el mejor algoritmo, utilizando técnicas automatizadas que maximizan su rendimiento sobre el conjunto de prueba.

## Comparación de Modelos

Una vez obtenidos los datos divididos en conjunto de entrenamiento y conjunto de prueba, el siguiente paso consistió en seleccionar los algoritmos de aprendizaje automático que se iban a evaluar para determinar cuál ofrecía mejor desempeño en este conjunto de datos analizado.

Para el modelado y caracterización de la variable valor agregado, se seleccionaron cuatro algoritmos de aprendizaje supervisado. Esta selección responde a la necesidad de contrastar diferentes arquitecturas lógicas para determinar cuál se adapta mejor a la complejidad de los microdatos de la EAM. Se eligieron algoritmos con distintos niveles de complejidad, basados en la revisión previa realizada en el marco teórico del presente proyecto. Los algoritmos que se implementaron son los siguientes (Bonaccorso, 2017):

- Regresión Lineal (Modelo Base): Utilizada como línea base para entender la relación directa entre los insumos y el producto generado por los establecimientos. Se

implementó como el modelo de referencia debido a su naturaleza paramétrica y alta interpretabilidad. Este algoritmo asume que el valor agregado se comporta como una función lineal de las variables de entrada (producción, consumo, salarios e inversión). Su inclusión es fundamental para validar si las relaciones económicas del sector pueden simplificarse a una tendencia recta o si, por el contrario, requieren modelos de mayor flexibilidad.

- K-Nearest Neighbors (K-NN): Este algoritmo se seleccionó por su enfoque no paramétrico basado en la proximidad. El K-NN estima el valor agregado identificando establecimientos con características operativas similares (vecinos cercanos) en un espacio n-dimensional. Es de gran utilidad para capturar patrones locales y observar si la similitud en la estructura de costos y personal se traduce en una generación de riqueza neta equivalente. Este modelo busca empresas similares en la base de datos para predecir el valor de una nueva.
- Árbol de decisión: Este modelo es fácil de interpretar y capaz de manejar variables categóricas y numéricas. Modela relaciones no lineales y jerárquicas. Se incluyó por su capacidad para modelar relaciones no lineales mediante reglas de decisión jerárquicas e intuitivas. A diferencia de la regresión lineal, el árbol de decisión puede segmentar automáticamente los datos según umbrales específicos en las variables (por ejemplo, niveles de producción bruta), permitiendo entender la estructura lógica de las empresas manufactureras de forma visual y directa.
- Random Forest: Es un modelo de ensamble que permite capturar la heterogeneidad de los datos, lo cual es esencial por las características propias de los datos de la industria manufacturera. Se seleccionó como el algoritmo de mayor complejidad y robustez. Al ser un método de ensamble que combina múltiples árboles de decisión mediante la técnica de *bagging*, reduce significativamente la varianza y el riesgo de sobreajuste (*overfitting*) que

presentan los árboles individuales. Es el modelo idóneo para manejar la alta heterogeneidad y los valores extremos identificados en la EAM, garantizando estimaciones más estables y una mayor capacidad de generalización.

**Conformación del Conjunto de Entrenamiento y Prueba**

Se aplicó una división del conjunto de datos bajo la regla 80/20. El 80% de los datos se destinó al entrenamiento de los algoritmos, mientras que el 20% de datos restantes se destinó para la validación de este entrenamiento. Este procedimiento garantiza que la evaluación del modelo represente su capacidad real de predecir datos nuevos y no simplemente su capacidad de memorizar la base de datos original.

1. Consistencia: Los datos ya no tienen escalas diferentes (millones vs. unidades).
2. Representatividad: Al limpiar outliers, el modelo representará el comportamiento industrial real de la mayoría de los establecimientos.

**Entrenamiento y Optimización de Modelos**

En esta etapa, cada algoritmo seleccionado fue sometido a un proceso de entrenamiento a partir del cual se generaron las métricas estadísticas. Tras finalizar el entrenamiento, se procedió a la comparación de los modelos utilizando el conjunto de prueba. A través de métricas como el $R^2$, MAE, RMSE y MAPE, se contrastó el desempeño de los modelos ejecutados, permitiendo seleccionar el mejor modelo como la herramienta con mayor precisión para la tarea de estimación.

Para el caso del modelo Random Forest, se ajustaron variables críticas como el número de estimadores y la profundidad máxima de los árboles mediante una búsqueda de hiperparámetros. Este proceso se orientó a maximizar el coeficiente de determinación ($R^2$),

asegurando que el modelo capturara la mayor proporción posible de la varianza del valor agregado y garantizando así una estimación más robusta y precisa.

## Evaluación del Modelo

### Métricas Estadísticas

La selección de las métricas estadísticas de evaluación para este proyecto se realiza basada estrictamente en la naturaleza de la variable objetivo: el valor agregado. Al tratarse de una variable cuantitativa continua (expresada en unidades monetarias), el problema analítico se define como una regresión y no como una clasificación.

Para validar la utilidad del modelo frente a los datos analizados se utilizaron las siguientes métricas (Ramasubramanian & Moolayil, 2019):

- Coeficiente de determinación ($R^2$): Seleccionado para medir la proporción de la varianza del valor agregado que el modelo logra capturar. Indica qué porcentaje de la variabilidad del valor agregado es explicada por el modelo. Un valor cercano a 1 indica una precisión alta.
- MAE (Error Medio Absoluto): Representa el promedio de cuánto se equivoca el modelo, se expresa en la misma escala que los datos originales; es decir, en este caso se expresa en términos monetarios (pesos colombianos) facilitando la interpretación financiera del margen de error promedio por establecimiento.
- RMSE (Root Mean Squared Error): Es la raíz del error cuadrático medio e indica fallos grandes, por lo cual permite entender que tan mal le va al modelo especialmente con grandes errores.
- MAPE (Error Porcentual Absoluto Medio): Mide el tamaño del error promedio en relación con los valores reales. En lugar de decir en cuántas unidades se falló (como el MAE), muestra que tan grande fue ese fallo respecto al tamaño de la empresa.

**Evaluación de las Métricas Estadísticas**

**Tabla 5**

*Comparación de Modelos*

| Modelo | $R^2$ | MAE ($) | RMSE ($) | MAPE (%) |
|---|---|---|---|---|
| Regresión Lineal | 1.00 | 0 | 0 | 0.00% |
| Random Forest (Optimizado) | 0.98 | 264,190 | 1,156,192 | 12.73% |
| Árbol de Decisión | 0.93 | 591,195 | 1,922,406 | 30.64% |
| K-Nearest Neighbors | 0.75 | 2,014,918 | 3,738,032 | 148.83% |

*Nota*. Resultado de las métricas implementadas.

A continuación, se presenta la interpretación de las métricas obtenidas en el proceso de comparación de los modelos para la caracterización de la variable valor agregado:

- Selección del modelo óptimo (Random Forest): El modelo de Random Forest se consolidó como la mejor herramienta de estimación con un coeficiente de determinación ($R^2$) de 0.98. Esto indica que el modelo logra explicar el 98% de la variabilidad del valor agregado a partir de las variables de consumo intermedio, salarios del personal permanente, activos fijos y producción bruta.

Su superioridad frente al Árbol de Decisión (0.93) y K-NN (0.75) confirma que el fenómeno de la productividad manufacturera requiere algoritmos de ensamble que capturen interacciones no lineales complejas entre los factores productivos.

- Regresión Lineal: En este modelo se observa un ajuste de $R^2 = 1.00$, si bien esto podría parecer ideal, en el análisis de datos industriales suele indicar una multicolinealidad perfecta (algunas variables de entrada suman directamente el valor de salida). Por tanto, se desestimó este modelo para evitar el sobreajuste (*overfitting*), priorizando la capacidad de generalización y robustez del modelo Random Forest.

- Árbol de Decisión: Logra un buen ajuste (0.93) pero su MAPE (30.64%) es casi el triple que el de Random Forest, evidenciando mayor inestabilidad.
- K-Nearest neighbors (KNN): Con un MAPE del 148.83%, queda totalmente descartado, demostrando que los modelos basados en cercanía no son aptos para la complejidad y escala de los datos de la EAM.
- Interpretación de los errores (MAE y MAPE): Respecto a la métrica MAE (Error Medio Absoluto) en el modelo Random Forest presenta una desviación promedio de $264,190 siendo el valor más bajo presentado por todos los modelos. En el contexto de los volúmenes financieros de la EAM, este margen de error es reducido. Respecto a la métrica MAPE (Error Porcentual Absoluto Medio) se puede concluir que es la métrica más intuitiva para la gestión. El valor de 12.73% obtenido por el modelo Random Forest significa que, en promedio, las estimaciones del modelo solo se alejan un 12.7% del valor real reportado por los establecimientos. Para un modelo basado en datos con alta variabilidad, un error cercano al 10% se considera de alta precisión.

Se debe tener en cuenta que el modelo Random Forest presenta un error del 12,73% y no más bajo ya que el sector manufacturero presenta una altísima heterogeneidad natural. Un MAPE del 12.73% es un resultado sobresaliente, ya que indica que el modelo ha logrado capturar el patrón de productividad de la mayoría de las empresas, logrando una herramienta de analítica confiable para la comprensión de las variables que impactan la variable valor agregado.

La implementación de los modelos seleccionados se diseñó con el objetivo de garantizar una comparación justa entre algoritmos bajo condiciones homogéneas. Para este proyecto aplicado se utilizó el conjunto de datos consolidado y depurado compuesto por 11,451 registros, obtenidos de los archivos históricos de la base de datos de la EAM del año 2023 y de la EAM del

año 2024. Cada modelo seleccionado fue entrenado utilizando exclusivamente la división en el conjunto de entrenamiento y evaluado sobre el conjunto de prueba.

Al comparar los modelos seleccionados y ejecutados en la sección anterior vemos los siguientes resultados en su desempeño:

- $R^2$ (Coeficiente de Determinación): Esta métrica nos permite comprender la capacidad explicativa. Es la capacidad predictiva sobre la variabilidad que el modelo logra entender. El resultado de 0.98 muestra que el Random Forest explica el 98% de la variación del valor agregado, un desempeño sobresaliente para datos económicos. La superioridad del modelo de Regresión Lineal en esta métrica se da ya que al tener un $R^2$ de 1 permite entender que existe una multicolinealidad perfecta (algunas variables de entrada suman directamente el valor de salida). Por tanto, se desestimó este modelo para evitar el sobreajuste, priorizando la capacidad de generalización y robustez del modelo Random Forest.

- MAE (Mean Absolute Error: Error Medio Absoluto): Esta métrica nos muestra el error promedio en términos absolutos. Es la diferencia promedio (en pesos) entre lo que el modelo predijo y lo que realmente pasó. El resultado del modelo Random Forest muestra que, en promedio, las predicciones del modelo se desvían $264.190 pesos de los valores reales, siendo el modelo con el menor valor en esta métrica, lo que muestra al Random Forest como el algoritmo con el error típico más reducido y robusto ante valores extremos. Es el error típico que se podría esperar en una predicción normal.

- RMSE (Root Mean Squared Error: Raíz del Error Cuadrático Medio): Esta métrica es el detector de grandes fallos. A diferencia del MAE, el RMSE penaliza mucho más los errores grandes. El valor $1,156,192 del modelo Random Forest es significativamente superior al MAE, esto nos indica que el modelo tiene dificultades con algunas empresas específicas

(posiblemente las muy grandes), siendo el modelo con el menor valor en esta métrica. Es vital para asegurar que el modelo no esté cometiendo errores graves en las empresas que más aportan a la producción industrial.

- MAPE (Mean Absolute Percentage Error: Error Porcentual Absoluto Medio): Esta métrica nos permite entender la precisión relativa. Es el error expresado en porcentaje. El resultado 12.73% del modelo Random Forest aunque parece alto, en contextos de datos con alta dispersión (como se vio en el coeficiente de variación de 277% de la caracterización inicial), es un resultado esperado. Indica que, en promedio, la predicción se desvía un 12% del valor real. Esta métrica revela que, si bien el modelo es preciso en términos monetarios para grandes empresas, las menores variaciones en empresas pequeñas generan desviaciones porcentuales altas, una característica natural de la heterogeneidad de la EAM.

Los resultados obtenidos permiten concluir que el algoritmo Random Forest es el modelo con mayor capacidad predictiva, logrando un equilibrio superior entre la precisión global ($R^2$ de 0.98) y el control de errores promedio. Mientras que la Regresión Lineal demostró una dependencia funcional (multicolinealidad) que invalida su uso práctico, el Random Forest logró capturar la esencia de los datos con un error MAE de $264,190, el más bajo de la comparativa. Esto confirma que los algoritmos de ensamble son la herramienta correcta para modelar la complejidad y heterogeneidad de la industria manufacturera colombiana, donde las relaciones entre variables no son lineales.

**Selección del Modelo Ganador**

A continuación, se presentan las conclusiones de los modelos evaluados (James et al., 2023):

- Modelo ganador: Random Forest: Con un $R^2$ de 0.98, este modelo es el más robusto ya que explica el 98% de la variación de la variable valor agregado. Además, tiene el MAE más bajo, lo que significa que es el modelo que tiene la mejor efectividad al predecir el dinero generado por las empresas.
- Segundo lugar: Árbol de Decisión: Este modelo logra un $R^2$ de 0.93. Tiene un desempeño muy respetable. Esto dice que la industria manufacturera se puede entender bien mediante reglas de decisión, por ejemplo, si se tiene más de 50 empleados y es del sector alimentos, entonces su valor es X. Sin embargo, al ser un solo árbol, pierde casi un 8% de precisión frente al Random Forest, que utiliza cientos de árboles trabajando en equipo.
- Tercer lugar: K-Nearest Neighbors (K-NN): La precisión de este modelo es del 75%. Este modelo intenta predecir el valor de una empresa basándose en las 5 empresas más similares que tiene cerca. Aunque es aceptable, su error MAE es más del doble que el del Random Forest, lo que indica que en la industria colombiana ser similar no siempre garantiza resultados económicos idénticos.
- El modelo base: Regresión Lineal: En este modelo se observa un $R^2$ de 1.00. Si bien esto podría parecer ideal, en el análisis de datos industriales indica una multicolinealidad perfecta (algunas variables de entrada suman directamente el valor de salida). Por tanto, se descartó este modelo para evitar el sobreajuste (*overfitting*), priorizando la capacidad de generalización y robustez del modelo Random Forest.

La Regresión Lineal asume que la industria se comporta de forma recta y predecible. Sin embargo, como vimos en la caracterización de las variables seleccionadas, la industria colombiana es muy heterogénea. El Random Forest gana porque es capaz de capturar relaciones no lineales, lo que lo hace mucho más apto para este tipo de microdatos del DANE.

La comparativa demuestra una evolución clara en la capacidad de aprendizaje: a medida que pasamos de modelos lineales a modelos de ensamble como Random Forest, el error medio se reduce y la precisión aumenta. Esto prueba que la complejidad de la industria manufacturera colombiana requiere algoritmos capaces de capturar patrones no lineales.

**Análisis de Importancia de las Variables**

Tras el entrenamiento y optimización del modelo Random Forest, se realizó un análisis de importancia de características para identificar qué variables tienen mayor peso en la estimación del valor agregado. Los resultados revelan una jerarquía clara en los determinantes de la productividad, por lo cual a continuación se muestra la importancia de cada variable predictora:

**Tabla 6**

*Importancia de Variables*

| Variable | Importancia |
|---|---|
| Prod_Bruta | 0.86 |
| Consumo_Interm | 0.11 |
| Sueldos_Pers_Perm | 0.03 |
| Inv_Activos | 000 |

*Nota*. Importancia de las variables predictoras.

- Producción bruta (86%): Esta es la variable dominante con una importancia del 0.86. Este resultado es teóricamente consistente, ya que la producción es el componente principal para el cálculo de la riqueza generada. El modelo identifica que las variaciones en el valor agregado están explicadas casi en su totalidad por la capacidad de producción de los establecimientos.

- Consumo intermedio (11%): Con un peso del 0.11, se posiciona como el segundo predictor más relevante. Esto confirma que el modelo ha capturado la relación del proceso de transformación industrial, ya que el valor agregado no solo depende de lo que se vende (producción), sino de la eficiencia con la que se utilizan los insumos.
- Variables de personal permanente contratado e inversión en activos fijos (3% y 0%): Los sueldos del personal permanente y la inversión en activos fijos muestran una importancia significativamente menor (0.03 y 0.00 respectivamente). Esto no significa que no sean importantes para una empresa, sino que, para el modelo predictivo, la información necesaria para estimar el valor agregado ya está contenida mayoritariamente en la variable de producción bruta y consumo intermedio. Respecto a la variable inversión en activos fijos se obtuvo un resultado de 0.00 el cual puede deberse a que la inversión es un dato muy estático o muchos establecimientos no reportaron información en esta variable. Podemos concluir que la inversión en activos fijos no es un predictor de corto plazo para el valor agregado.

A partir del anterior resultado se puede concluir que, para tener un modelo predictivo altamente preciso en la EAM, basta con monitorear de cerca la producción bruta y el consumo intermedio, simplificando la futura recolección de datos.

**Generación de Combinaciones**

Si bien ya sabemos que el Random Forest es el mejor modelo, ahora el objetivo es encontrar la configuración exacta de sus piezas internas para extraerle hasta el último punto de precisión posible. En este proceso de configuración se probaron diferentes opciones. Las más importantes para este proyecto aplicado de la EAM son:

- *n_estimators:* Definir cuál es la mejor cantidad de árboles a usar. Más árboles suelen dar más estabilidad, pero tardan más. Significa que el modelo final utiliza un ensamble de

100 árboles de decisión trabajando en paralelo. Es el equilibrio justo entre potencia de cómputo y precisión.

- *max_depth:* Definir la profundidad de los árboles. Si son muy profundos, pueden memorizar empresas específicas; si son muy cortos, no aprenden bien. Indica que, para los datos de la EAM, es mejor dejar que los árboles crezcan hasta que todas las hojas sean puras. El modelo tiene suficiente información para ser profundo sin caer en el error de memorizar, gracias a la cantidad de datos que se tienen.
- *min_samples_split:* Definir cuántos datos necesita un nodo para dividirse. Ayuda a controlar que el modelo sea general y no solo para un caso. Significa que el modelo es capaz de encontrar patrones muy específicos incluso en grupos pequeños de empresas.

Ejecutamos una técnica llamada *GridSearchCV* con validación cruzada. Esto significa que el código probará múltiples combinaciones de configuraciones y verificará cuál funciona mejor no solo una vez, sino en diferentes partes de los datos para asegurar que el resultado sea estable. Esta herramienta prueba automáticamente todas las combinaciones que se le dan al modelo y genera la mejor combinación.

- Optimización final: Se podría incrementar el valor de $R^2$ en pequeñas cantidades. En proyectos de alto nivel, cada decimal cuenta.
- Robustez: Al usar validación cruzada (cv=3), aseguramos que el modelo funcione bien con cualquier parte de la base de datos, no solo con una porción.
- Eliminación del sesgo: Al usar cv=3 (Validación cruzada), se demuestra que el $R^2$ no es producto de la suerte con un grupo de datos específico, sino que el modelo es consistente en toda la base de datos de la EAM.

- Prevención del sobreajuste: Al probar diferentes *max_depth* y *min_samples_split*, nos aseguramos de que el modelo no esté memorizando las empresas, sino aprendiendo las reglas generales de la industria manufacturera.

Para maximizar la confiabilidad de las predicciones, se realizó un proceso de optimización de hiperparámetros mediante la técnica de *Grid Search* con validación cruzada. Se evaluaron 18 combinaciones distintas de parámetros estructurales del Random Forest. El resultado permitió estabilizar la precisión del modelo en un $R^2$=0,9739, garantizando que el algoritmo sea capaz de generalizar sus predicciones para pequeñas y grandes establecimientos industriales, minimizando el riesgo de sobreajuste.

**Figura 5**

*$R^2$ Optimizado*

```
--- OPTIMIZACIÓN COMPLETADA ---
Mejores parámetros encontrados: {'bootstrap': True, 'max_depth': None, 'min_samples_split': 2, 'n_estimators': 100}
Mejor R² (Precisión) tras optimizar: 0.9739
```

*Nota*. Resultado obtenido del proceso de optimización.

Con el objetivo de maximizar la capacidad predictiva del algoritmo Random Forest, se realizó un proceso de optimización de hiperparámetros mediante la técnica de búsqueda en cuadrícula (Grid Search) con validación cruzada. Este procedimiento permitió identificar la configuración óptima para el dataset de la EAM, resultando en un modelo compuesto por 100 estimadores y técnica de bootstrapping activada. Tras la optimización, el modelo alcanzó un $R^2$ de 0.9739, lo que representa una mejora significativa en la estabilidad de las estimaciones y asegura que el modelo no presenta sobreajuste (overfitting), sino una alta capacidad de generalización.

A partir del proceso de optimización se puede afirmar que de cada 100 pesos de valor agregado que genera una empresa, el modelo predice correctamente casi 98 de ellos basándose únicamente en los datos operativos.

Se debe tener en cuenta que para la ejecución de la validación cruzada este número no se obtuvo de una sola prueba, sino de 54 ajustes diferentes. Esto garantiza que el modelo es estable y no tuvo suerte con un grupo de datos al azar.

Tras realizar una búsqueda exhaustiva de hiperparámetros (*Grid Search*) con validación cruzada, se identificó la configuración óptima para el algoritmo de Random Forest. El proceso de optimización permitió elevar la precisión del modelo a un 97.3%. Esta fase garantiza que el modelo no presenta sobreajuste (*overfitting*) y que posea una capacidad de generalización robusta. Los parámetros finales confirman que la arquitectura del modelo aprovecha la profundidad total de los datos de la EAM para capturar las particularidades de los establecimientos analizados, consolidándose como una herramienta necesaria para comprender el valor agregado.

## Conclusiones

El análisis exploratorio de datos permitió confirmar que la industria manufacturera colombiana presenta una alta heterogeneidad estructural, evidenciada en el hecho de que existe una dispersión extrema en los datos analizados al tener coeficientes de variación superiores al 220% y un valor máximo del 1493% (Inversión en activos fijo). Debido a esta situación se implementaron técnicas de normalización y escalamiento de variables que permitieron que el modelo no se distorsionara ante las diferencias entre pequeñas y grandes empresas.

La fase de procesamiento permitió transformar la información heterogénea de la EAM en un conjunto de datos técnicamente robusto, reduciendo el ruido estadístico mediante la limpieza de valores atípicos; además, se realizó la implementación de técnicas de normalización de variables que fue determinante para equilibrar las dimensiones económicas de establecimientos de distintos tamaños. A partir de estas acciones se conformó un dataset con integridad y consistencia, lo que permitió que los algoritmos de aprendizaje fueran eficientes y evitaran sesgos derivados de la disparidad original de la muestra.

El proceso de limpieza y tratamiento de valores atípicos (outliers) resultó crítico para la estabilidad del modelo. La reducción del dataset original de 13,297 a 11,451 registros permitió eliminar el ruido estadístico causado por establecimientos con comportamientos atípicos, garantizando que el aprendizaje del modelo se centrara en los patrones representativos del sector. Asimismo, la optimización de hiperparámetros mediante Grid Search fue el factor determinante para alcanzar una precisión del 97.39% en el conjunto de prueba, minimizando el riesgo de sobreajuste.

A través del análisis de importancia de características del modelo Random Forest, se determinó que la producción bruta y el consumo intermedio son los predictores determinantes en

la estimación del valor agregado, concentrando entre ambos el 97% del peso explicativo del modelo. Este hallazgo valida que la generación de valor agregado en el sector manufacturero colombiano está vinculada primordialmente a la dinámica de transformación de insumos en productos terminados.

Por otro lado, se observó que variables como los sueldos del personal permanente y la inversión en activos fijos poseen una influencia marginal en la variabilidad del valor agregado. Esto sugiere que, si bien el capital humano y la infraestructura son pilares estructurales, la generación de valor agregado depende de la eficiencia operativa y el volumen de producción. Estos resultados permiten identificar que las intervenciones orientadas a optimizar el uso de insumos y maximizar la producción bruta tendrán un impacto directo y predecible en el fortalecimiento del valor agregado.

Se concluye que el algoritmo Random Forest es la arquitectura más robusta para la caracterización del valor agregado en la industria manufacturera colombiana, alcanzando un R2 de 0.98 y un MAE de $264,190. La superioridad de este modelo de ensamble sobre la Regresión Lineal y el K-NN demuestra que la generación de valor no es lineal y presenta una alta heterogeneidad que solo puede ser capturada mediante algoritmos capaces de segmentar los datos y promediar múltiples estimadores.

Se determinó que la relación entre los insumos consumidos y la generación de valor en la industria manufacturera colombiana es de naturaleza no lineal y altamente compleja, porque la Regresión Lineal mostró una multicolinealidad perfecta (algunas variables de entrada suman directamente el valor de salida), lo que llevo a que se desestimara este modelo para evitar el sobreajuste (overfitting), priorizando la capacidad de generalización y robustez del modelo

Random Forest. Este hallazgo demuestra que los métodos estadísticos tradicionales son insuficientes para capturar la heterogeneidad del sector manufacturero.

Se validó la robustez del modelo Random Forest, el cual, a pesar de la heterogeneidad estructural del sector, presentó un Error Medio Absoluto (MAE) de $264,190. Esta capacidad predictiva mitiga las distorsiones causadas por la falta de uniformidad en los reportes de los establecimientos industriales, consolidando una herramienta de modelado y comprensión de las variables del sector manufacturero.

La EAM es una fuente estratégica de información de vital importancia para el análisis del sector manufacturero de Colombia ya que los datos de esta operación estadística ofrecen un nivel de detalle que permite ir más allá de los análisis descriptivos tradicionales, facilitando el diseño de modelos predictivos orientados a comprender las dinámicas productivas del sector industrial y de esta forma entender el funcionamiento del sector manufacturero en Colombia.

El uso de métodos de aprendizaje automático permite identificar patrones no evidentes mediante técnicas convencionales, fortaleciendo la capacidad del análisis para entender las razones de los comportamientos empresariales y apoyar la toma de decisiones basadas en evidencia.

## Recomendaciones

Dado que el modelo actual utiliza datos de dos años se podría integrar información de múltiples períodos anuales, lo que permitiría capturar efectos temporales que el modelo actual no tiene en cuenta. Para profundizar los resultados obtenidos también se pueden incorporar variables exógenas como la tasa de cambio (TRM) y el índice de precios al productor (IPP), lo que permitiría entender si la volatilidad macroeconómica afecta la capacidad de generar valor agregado en las distintas actividades de manufactura.

La Encuesta Anual Manufacturera es una operación estadística que recolecta información de diversas variables, por lo cual se podrían incluir en el modelo otras variables que pueden permitir entender de manera correcta la generación de valor agregado; tales como, personal temporal contratado, ventas al exterior o compras en el exterior.

El hallazgo de una dispersión extrema en variables críticas muestra la necesidad de implementar mecanismos de validación de datos en tiempo real durante la captura de la información. Se propone que el DANE integre algoritmos de detección de anomalías en los formularios digitales. Esto reduciría la presencia de errores de digitación o reportes inconsistentes que generan coeficientes de variación altos, mejorando la calidad de los insumos para futuros modelos de *Machine Learning*.

Se recomienda realizar una segmentación previa de la muestra (*Clustering*) por tamaño de empresa antes del entrenamiento. Entrenar modelos específicos para cada clúster podría mejorar significativamente la precisión, ya que las dinámicas operativas de una microempresa no responden a los mismos patrones que las de una gran empresa manufacturera.

El modelo puede automatizar la detección de errores de digitación o reportes inconsistentes. Al recibir un nuevo formulario, el sistema ingresa los datos al modelo. Por

ejemplo, si el valor agregado real se aleja más de 2 desviaciones estándar del predicho, el registro se marca automáticamente para revisión, y de esta forma enfocarse en aquellos registros que el modelo identifica como sospechosos o ineficientes.

Se pueden generar reportes automáticos por sector manufacturero ya que el modelo permite comparar empresas similares para identificar quiénes están quedándose atrás a pesar de tener los mismos recursos que sus competidores. Si una empresa tiene un consumo intermedio alto pero su valor agregado está por debajo de la norma aprendida por el Random Forest, se le puede priorizar para programas de asistencia técnica.

La EAM es una operación estadística de periodicidad anual, por lo cual no se dispone de información de períodos de estudio cortos que permitan capturar variaciones en el corto plazo, por lo anterior se podría realizar la incorporación de la información recolectada en la operación estadística Encuesta Mensual Manufacturera con Enfoque Territorial, la cual tiene una periodicidad mensual y recolecta información de algunos establecimientos que también reportan información en la Encuesta Anual Manufacturera.

El DANE podría fortalecer la EAM mediante nuevas variables relacionadas con innovación, digitalización y adopción tecnológica. Por esto se sugiere ampliar la información recolectada de estas variables que estén directamente relacionadas con los procesos de investigación y desarrollo y con el proceso de producción con el fin de enriquecer las posibilidades analíticas.

Se puede adaptar e implementar este proyecto a otras operaciones estadísticas en Colombia como la Encuesta Anual de Servicios y Encuesta Anual de Comercio y de esta forma ejecutar un seguimiento de la actividad de servicios y comercio en Colombia.

La visualización es clave para la comunicación de resultados, por lo que se recomienda integrar el modelo en una interfaz interactiva actualizada y accesible para los usuarios de información que realizan la toma de decisiones. La elaboración de gráficos interactivos facilitará la comunicación de los hallazgos y su integración en la planeación estratégica del sector manufacturero.

Se recomienda generar un sistema de alerta temprana, para lo cual se podría modificar el modelo de regresión hacia uno de clasificación binaria que, al detectar caídas críticas en el valor agregado proyectado emita alertas de riesgo.

## Referencias Bibliográficas

Ahumada, D. P. (2023). *Técnicas de minería de datos para el análisis de pruebas SABER.* [Tesis de Maestría, Universidad Nacional de Colombia]. Repositorio Institucional. https://repositorio.unal.edu.co/handle/unal/84182

Arango, M. (2021). *Scikit-forecasts: Una librería en Python para el pronóstico de series de tiempo no lineales.* [Tesis de Maestría, Universidad. Nacional de Colombia]. Repositorio Universidad Nacional de Colombia. https://repositorio.unal.edu.co/handle/unal/81979

Araque, G., & Giampietro, V. (2023). El Big Data aplicado en la industria 4.0: un caso en el sector textil colombiano con un enfoque en la inteligencia de negocios. *Cuaderno Activa*, *14*(1). https://doi.org/10.53995/20278101.1176

Bonaccorso, G. (2017). *Machine learning algorithms: Reference guide for popular algorithms for data science and machine learning*. Packt Publishing.

Buitrago, N. S. (2021). *Análisis de los precios del metro cuadrado de la vivienda nueva en la ciudad de Bogotá.* [Tesis de Maestría, Universidad Nacional de Colombia]. *Repositorio de la Universidad Nacional de Colombia.* https://repositorio.unal.edu.co/handle/unal/79630

Carmona, M., Carvajal, Y., Aguirre, S. M., Ocampo, F. J., & Flórez, A. M. (2020). Determinantes del crecimiento empresarial en el sector manufacturero colombiano. *Panorama Económico*, *28*(1), 1–15. https://doi.org/10.32997/pe-2020-2665

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1. Step-by-step data mining guide*. DaimlerChrysler.

Confecámaras. (2023). *La supervivencia empresarial en Colombia* (Confecámaras, Ed.).

Departamento Administrativo Nacional de Estadística (DANE). (2025a). *Encuesta Anual Manufacturera (EAM)*. https://www.dane.gov.co/index.php/estadisticas-por-tema/industria/encuesta-anual-manufacturera-enam

Departamento Administrativo Nacional de Estadística (DANE). (2025b). *Ficha Metodológica Encuesta Anual Manufacturera-EAM*.

Departamento Administrativo Nacional de Estadística (DANE). (2025c). *Metodología General Encuesta Anual Manufacturera-EAM*.

Devore, J. L. (2018). *Fundamentos de Probabilidad y Estadística*. Cengage.

Fierro, C., Castillo, V., & Torres, C. (2022). Análisis comparativo de modelos tradicionales y modernos para pronóstico de la demanda: enfoques y características. *RIDE Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, *12*(24). https://doi.org/10.23913/ride.v12i24.1203

Gao, H., Kou, G., Liang, H., Zhang, H., Chao, X., Li, C. C., & Dong, Y. (2024). Machine learning in business and finance: a literature review and research opportunities. *Financial Innovation*, *10*(1). https://doi.org/10.1186/s40854-024-00629-z

Gerón, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems*. O´Reilly.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer.

Hazelton, M. L. (2010). Univariate Linear Regression. *International Encyclopedia of Education, Third Edition*, 482–488. https://doi.org/10.1016/B978-0-08-044894-7.01373-7

Huber, G., Gürtler, L., & Gento, S. (2017). La aportación de la estadística exploratoria al análisis de datos cualitativos. *Perspectiva Educacional*, *57*(1), 50–69. https://doi.org/10.4151/07189729-Vol.57-Iss.1-Art.611

IBM. (2020). *Guía de CRISP-DM de IBM SPSS Modeler*.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning with Applications in Python*. Springer.

Kelleher, J., & Tierny, B. (2022). *Ciencia de Datos*. Ediciones UC.

López, E., Cáceres, A., Grillo, S., & Herrera, E. (2025). Evaluación de exactitud de Naive Bayes y Regresión Logística para clasificación con atributos y clases binarios. *Reportes Científicos de La FACEN*, *13*(1), 73–84. https://doi.org/10.18004/rcfacen.2022.13.1.73

Mariño, J. A. (2023). *Una comparación entre modelos estadísticos y de Machine Learning para la predicción de series de tiempo multivariadas.* [Tesis de Maestría. Universidad Nacional de Colombia]. Repositorio Universidad Nacional de Colombia. https://repositorio.unal.edu.co/handle/unal/84522

Massaron, Luca, & Boschetti, Alberto. (2016). *Regression analysis with Python: Learn the art of regression analysis with Python*. Packt Publishing.

OECD. (2021). *Business Insights on Emerging Markets 2021*. https://www.oecd.org/dev/

Parra, J. (2002). Análisis exploratorio y análisis confirmatorio de datos. *Espacio Abierto*, *11*(1). https://www.redalyc.org/articulo.oa?id=12211106

Pramanik, J., Samal, A. K., Sahoo, K., & Pani, S. (2019). Exploratory Data Analysis using Python. *International Journal of Innovative Technology and Exploring Engineering*, *8*, 4727–4735.

Ramasubramanian, K., & Moolayil, J. (2019). *Applied Supervised Learning with R*. Packt Publishing, Limited.

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/J.PROCS.2021.01.199

Silva, A. M. (2025). *Análisis de los factores que contribuyen al cierre temprano de las pymes en Colombia antes de su consolidación.* [Tesis de Maestría. Universidad Nacional Abierta y a Distancia]. Repositorio Universidad Nacional Abierta y a Distancia. https://repository.unad.edu.co/handle/10596/70703

Tascón, M. T., & Castaño, F. J. (2012). Variables y Modelos Para La Identificación y Predicción Del Fracaso Empresarial: Revisión de La Investigación Empírica Reciente. *Revista de Contabilidad*, *15*(1), 7–58. https://doi.org/10.1016/S1138-4891(12)70037-7

Triola, M. (2004). *Estadística*. Pearson.

Wirth, R., & Hipp, J. (2000). *Crisp-DM: Towards a standard process model for data mining*. http://hdl.handle.net/20.500.12010/36985