



Determinantes del Valor Bruto de Producción en la Industria Manufacturera: Un Estudio Basado en Machine Learning

Martha Beatriz Flores Romero

Facultad de Contaduría y Ciencias Administrativas
Universidad Michoacana de San Nicolás de Hidalgo
martha.flores@umich.mx

Jennifer López Chacón

Facultad de Biología
Universidad Michoacana de San Nicolás de Hidalgo
jennifer.lopez@umich.mx

Alfredo Raya Montaña

Instituto de Física y Matemáticas
Universidad Michoacana de San Nicolás de Hidalgo
alfredo.raya@umich.mx

Resumen:

Este estudio examina el Valor Bruto de Producción (VBP) en la industria manufacturera de México, su evolución histórica y su impacto en el crecimiento económico. Se analiza la inversión pública, la política de sustitución de importaciones y las crisis económicas que han afectado al sector. La investigación emplea Machine Learning para identificar los principales factores que influyen en la producción manufacturera, utilizando modelos como Random Forest y XGBoost. El análisis destaca que la inversión en equipo de transporte, estructuras residenciales y construcción no residencial son determinantes clave del crecimiento manufacturero. La metodología SHAP permitió interpretar la influencia de cada variable en el modelo. Random Forest obtuvo el mejor desempeño con un R^2 de 0.9922, demostrando su precisión predictiva. Se concluye que el Machine Learning es una herramienta útil para optimizar la inversión en la industria manufacturera y mejorar la toma de decisiones económicas.

Palabras Claves: Manufactura; Industrialización; Machine learning, SHAP.

Códigos JEL: L60; L16; C45; C52.



Determinants of Gross Production Value in the Manufacturing Industry: A Study Based on Machine Learning

Abstract:

This study examines the Gross Production Value (GVP) in Mexico's manufacturing industry, its historical evolution, and its impact on economic growth. Public investment, import substitution policy, and economic crises that have affected the sector are analyzed. The research uses Machine Learning to identify the main factors that influence manufacturing production, using models such as Random Forest and XGBoost. The analysis highlights that investment in transportation equipment, residential structures, and non-residential construction are key determinants of manufacturing growth. The SHAP methodology allowed the interpretation of the influence of each variable in the model. Random Forest obtained the best performance with an R^2 of 0.9922, demonstrating its predictive accuracy. It is concluded that Machine Learning is a useful tool to optimize investment in the manufacturing industry and improve economic decision-making.

Keywords: Manufacturing; Industrialization; Machine learning, SHAP.

JEL Codes: L60; L16; C45; C52.



1. Introducción

El Valor Bruto de Producción (VBP) en la industria manufacturera es un indicador clave para medir la actividad económica y el crecimiento de un país, reflejando el valor total de los bienes y servicios producidos por este sector. Este sector es considerado uno de los principales motores de crecimiento económico en muchas economías, especialmente en países en desarrollo como México, donde la manufactura representa una proporción significativa del Producto Interno Bruto (PIB) (OECD, s.f.). La industria manufacturera es responsable no solo de la generación de empleo, sino también de la innovación tecnológica y de la competitividad global, lo que la convierte en un área crucial para el análisis económico (OECD, s.f.).

El uso de técnicas avanzadas de Machine Learning (ML) para el análisis del VBP ha ganado relevancia en la investigación económica, ya que permite identificar patrones complejos y no lineales en los datos que los métodos tradicionales no pueden captar. El Machine Learning se ha utilizado exitosamente en diversas áreas, como la predicción de la productividad, el análisis de la eficiencia en la producción y la identificación de los factores que afectan la rentabilidad del sector manufacturero (Lee & Lim, 2021). A través de modelos de predicción, como los árboles de decisión, redes neuronales y modelos de ensamble, se pueden obtener insights precisos sobre las variables que influyen en el VBP, lo que permite tomar decisiones más informadas sobre políticas económicas y empresariales (Zhang et. al., 2020).

Sin embargo, a pesar de la creciente aplicación de ML en este campo, la investigación sigue siendo incipiente, especialmente en contextos específicos de países en desarrollo. Este estudio tiene como objetivo utilizar técnicas de Machine Learning para analizar los determinantes del VBP en la industria manufacturera, centrándose en factores como la inversión en tecnología, la productividad laboral, las políticas públicas y las condiciones macroeconómicas. El análisis buscará proporcionar una mejor comprensión de cómo estas variables afectan la producción en este sector, con el fin de ofrecer recomendaciones para optimizar el desempeño de la manufactura en economías emergentes.



En este trabajo se presenta en el primer apartado la historia de la industria manufacturera de México y su estado actual. Posteriormente, se muestra cómo se han utilizado técnicas de aprendizaje automático en la industria manufacturera. Luego, se presenta la metodología utilizada y los resultados obtenidos. Finalmente, se dan a conocer las conclusiones encontradas en este trabajo de investigación.

2. Industria Manufacturera en México

Después de la Segunda Guerra Mundial, México vivió una fase de crecimiento económico que incluyó un aumento significativo en la productividad en países industrializados, como Japón y los países occidentales. Entre 1948 y 1949, México experimentó una devaluación de su moneda de 4.85 a 8.65 pesos por dólar, alcanzando en 1954 un valor de 12.50 pesos por dólar (Banco Mundial, 2018). La tasa de cambio y el control de importaciones fueron cruciales para proteger la industria nacional.

En ese contexto, el gobierno mexicano implementó la tasa de cambio como un mecanismo proteccionista, aunque el Acuerdo de Comercio entre México y Estados Unidos (1943-1950) limitó la adopción plena de estas políticas (Tello, 1980). La inversión pública, que representó el 41.84% del total entre 1945 y 1959, fue fundamental en sectores estratégicos como electricidad, gas y petróleo (Cárdenas, 2010). Este periodo estuvo marcado por un modelo de crecimiento con devaluación e inflación, acompañado de un fuerte apoyo al sector agrícola, que fue crucial para el desarrollo industrial al proporcionar materias primas, mano de obra y alimentos (Villarreal, 1998).

Durante la fase avanzada de la Sustitución de Importaciones (1959-1970), México experimentó una mejora en la industrialización con un crecimiento promedio anual del 7.1%. El modelo estabilizador de crecimiento, que incluía un mayor control sobre las importaciones y el uso de aranceles, continuó siendo central en la estrategia económica del país. Para 1970, el control de importaciones alcanzó el 68%, un aumento respecto al 25% de 1956 (Cárdenas, 2010).



Sin embargo, la crisis petrolera de los años 70 afectó a la economía global, incluido México, reduciendo el crecimiento y alterando las políticas de desarrollo. A partir de 1970, el país enfrentó un lento crecimiento económico debido a la dependencia del petróleo y a un elevado nivel de deuda externa, que pasó de 19,602 millones de dólares en 1976 a 58,874 millones en 1982 (Aspe, 1993; Calva, 2019). Esto resultó en una devaluación de la moneda de 12.50 a 19.70 pesos por dólar en 1976 (Calva, 2019).

En el periodo de 1976-1981, a pesar de un crecimiento limitado y una alta inflación, no hubo una desindustrialización, aunque las importaciones del sector manufacturero aumentaron (Villarreal, 2013). A medida que la economía pasaba de un modelo industrial sustitutivo a un modelo petrolero exportador, la dependencia de los hidrocarburos se intensificó, con dos terceras partes de las exportaciones provenientes del petróleo en 1981 (Fajnzylber, 1990).

La crisis de 1982, marcada por una devaluación del 500% de la moneda (de 25 a 150 pesos por dólar), significó un estancamiento económico. Durante este tiempo, el gobierno de Miguel de la Madrid implementó el Programa Inmediato de Reordenación Económica (PIRE), con el objetivo de controlar la inflación y estabilizar la economía (Banco Mundial, 2018). La década de 1980 fue conocida como la "década perdida", con una estrategia económica de exportación de manufacturas que sustituyó gradualmente las importaciones, aunque con limitaciones debido a la concentración en unos pocos productos y mercados (Fajnzylber, 1990).

En la década de 1990, el gobierno de Carlos Salinas de Gortari promovió la apertura económica, la privatización y la liberalización del comercio, destacándose la firma del Tratado de Libre Comercio de América del Norte (TLCAN). Aunque las exportaciones de manufacturas aumentaron, la dependencia del mercado estadounidense continuó siendo una preocupación central para la economía mexicana (Salinas, 2000).



3. Modelos de Aprendizaje Automático

Los modelos de aprendizaje automático (o machine learning, ML) son técnicas computacionales que permiten a los sistemas aprender patrones a partir de datos, sin ser programados explícitamente, para realizar tareas como predicción, clasificación o optimización. Estos modelos se basan en algoritmos que identifican relaciones complejas en los datos, lo que los hace especialmente útiles en aplicaciones donde las reglas tradicionales son difíciles de definir. En el contexto de la industria manufacturera, el aprendizaje automático ha revolucionado áreas como la predicción de fallas, la optimización de procesos y la gestión de la producción. Trabajos pioneros, como el de Mitchell (Mitchell (1997)), permitieron definir el aprendizaje automático como la capacidad de un sistema para mejorar su desempeño mediante la experiencia, por lo que sentaron las bases teóricas para su aplicación práctica. Más recientemente, nuevos enfoques (Goodfellow et al. (2016)) han avanzado en el campo del aprendizaje profundo (deep learning), demostrando cómo las redes neuronales pueden manejar grandes volúmenes de datos y tareas complejas, como el análisis de imágenes y la predicción de series temporales.

En la industria manufacturera, diversos modelos de aprendizaje automático han sido aplicados para abordar desafíos específicos, cada uno con ventajas particulares según el contexto. Los modelos de regresión lineal y logística son ampliamente utilizados para predecir variables continuas y categóricas, como el valor bruto de producción (VBP) o la probabilidad de fallas en equipos, respectivamente (Wuest et al., 2016). Por otro lado, los árboles de decisión y los métodos de ensamble, como Random Forest y Gradient Boosting han demostrado ser efectivos para manejar datos no lineales y capturar interacciones complejas entre variables, lo que los hace ideales para la optimización de procesos y la clasificación de defectos (Kang et al., 2020). Además, las redes neuronales profundas (DNN) y las redes neuronales convolucionales (CNN) han ganado popularidad en aplicaciones avanzadas, como el monitoreo predictivo de maquinaria y la inspección automatizada de calidad, gracias a su capacidad para procesar grandes volúmenes de datos y aprender patrones intrincados (Zhang et al., 2020). Estos modelos, combinados con técnicas de preprocesamiento de datos y validación cruzada, permiten a las empresas manufactureras mejorar la precisión de sus predicciones y tomar decisiones más informadas.



3. Metodología

Fuente de Datos

Los datos utilizados en este estudio provienen de la base de datos LAKLEMS, un proyecto de cooperación técnica financiado por el Banco Interamericano de Desarrollo. Esta base de datos se basa en series estadísticas elaboradas por agencias estadísticas de los países participantes y ha sido desarrollada por la Universidad de Santiago de Chile con asesoramiento del Vienna Institute for International Economic Studies (Wiiw), siguiendo la metodología de EUKLEMS 2019 (LAKLEMS, 2021). Para este análisis, se utilizaron datos del sector manufacturero en el período 1990-2018, con todas las variables homologadas a moneda nacional y expresadas como índice de volumen, donde el año base es 2011=100. La base de datos se almacenó en Google Drive y se accedió a ella mediante Google Colab, una plataforma basada en la nube que permite ejecutar código en Python sin necesidad de instalación local.

Variables del Estudio

El análisis se centró en la predicción del Valor Bruto de Producción (GO_P) a partir de un conjunto de variables independientes que representan distintos tipos de inversión en activos físicos e intelectuales, así como servicios laborales:

Ip_IT: Equipos computacionales

Ip_CT: Equipos de comunicación

Ip_Soft_DB: Software

Ip_TraEq: Equipo de transporte

Ip_OMach: Otra maquinaria y equipos

Ip_OCon: Construcción no residencial

Ip_RStruc: Estructura residencial

Ip_Cult: Activos cultivables

Ip_OIPP: Otros activos de propiedad intelectual

LAB_QI: Servicios laborales



La variable dependiente es:

GO_P: Valor bruto de la producción

Preprocesamiento y División de Datos

El conjunto de datos se cargó en un DataFrame de pandas y se realizó una exploración inicial para identificar posibles valores faltantes y verificar la estructura de las columnas. Posteriormente, los datos se dividieron en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%) utilizando la función `train_test_split` de scikit-learn, garantizando la reproducibilidad mediante la semilla `random_state=42`.

Selección del Modelo y Entrenamiento

Se probaron inicialmente dos modelos de aprendizaje automático para la predicción de GO_P:

Random Forest Regressor: Un modelo basado en árboles de decisión que crea múltiples árboles de regresión y promedia sus predicciones para mejorar la precisión y reducir la varianza. Se configuró con `n_estimators=100` y `random_state=42`.

XGBoost Regressor: Un modelo de boosting basado en árboles que optimiza de manera secuencial el error de predicción. Se probó con `n_estimators=100`, `learning_rate=0.1` y `max_depth=3`.

Después de evaluar el error cuadrático medio (MSE) y el coeficiente de determinación (R^2) en ambos modelos, se decidió utilizar Random Forest Regressor, ya que proporcionó un mejor rendimiento en términos de precisión predictiva.

Evaluación del Modelo

Se evaluó el desempeño del modelo mediante dos métricas:

Error Cuadrático Medio (MSE): Mide la magnitud promedio del error en la predicción.

Coefficiente de Determinación (R^2): Indica qué porcentaje de la variabilidad de GO_P es explicada por las variables independientes.



Análisis de Importancia de Variables

Para interpretar los resultados, se calculó la importancia de cada variable en la predicción del modelo de Random Forest. Esto permitió identificar qué factores tienen mayor impacto en el valor bruto de la producción. Además, se utilizó SHAP (SHapley Additive exPlanations) para evaluar la influencia individual de cada variable en las predicciones. Se generaron los siguientes análisis visuales:

Gráfico de Importancia de Variables: Muestra las variables más influyentes en la predicción de GO_P.

Gráficos Summary Plot de SHAP: Ilustran el impacto de cada variable en la predicción del modelo.

SHAP Dependence Plot: Permite observar cómo cada variable interactúa con el modelo de predicción.

Implementación y Visualización

Todo el análisis se realizó en Google Colab, aprovechando su capacidad para trabajar con grandes volúmenes de datos y su integración con bibliotecas de Python como pandas, numpy, scikit-learn, xgboost, shap y matplotlib. Los resultados fueron visualizados mediante gráficos de barras para la importancia de variables y gráficos SHAP para una interpretación detallada de las predicciones.

4. Resultados

Evaluación del Desempeño del Modelo

Se evaluó el desempeño de los modelos de aprendizaje automático en la predicción del Valor Bruto de Producción (GO_P). Los resultados obtenidos fueron los siguientes:

Modelo	MSE	R²
Random Forest	7.5132	0.9922
XGBoost	49.3086	0.9487

El modelo de Random Forest obtuvo un menor error cuadrático medio y un coeficiente de determinación más alto en comparación con XGBoost, lo que indica que ofrece una mejor capacidad predictiva para este conjunto de datos.

Importancia de Variables

Se analizó la contribución de cada variable en el modelo de Random Forest. La distribución de esto se muestra en la Figura 1.

Figura 1. Gráfica correspondiente al análisis de la importancia de variables.



Fuente: Elaboración propia.

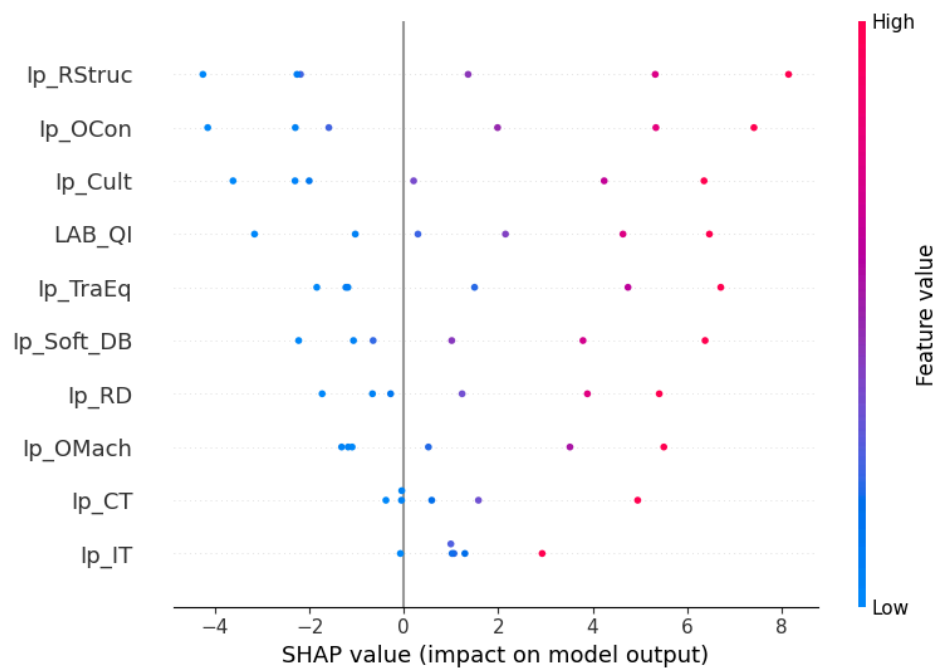
Los resultados indican que las variables con mayor impacto en la predicción del Valor Bruto de Producción (GO_P) están relacionadas principalmente con infraestructura y equipamiento. El equipo de transporte (Ip_TraEq), la estructura residencial (Ip_RStruc) y la construcción no residencial (Ip_OCon) se destacan como los factores más influyentes, seguidos por los servicios laborales (LAB_QI) y la inversión en otra maquinaria y equipos (Ip_OMach). Esto sugiere que la inversión en activos físicos y en la fuerza laboral juega un papel clave en el desempeño de la industria manufacturera. En contraste, la variable Ip_IT (Equipos computacionales) mostró un

impacto significativamente menor en la predicción, lo que podría indicar que su influencia en la producción no es tan directa o inmediata como la de otros factores más tangibles.

Análisis SHAP: Interpretabilidad del Modelo

Para una interpretación más detallada, se realizó un análisis utilizando SHAP (SHapley Additive exPlanations). Este método permite visualizar el impacto de cada variable en las predicciones del modelo.

Figura 2. Gráfica correspondiente al análisis SHAP.



Fuente: Elaboración propia.

Los resultados del análisis SHAP confirman que variables como Ip_RStruc, Ip_OCon y LAB_QI tienen una influencia significativa en el modelo. Los valores más altos de estas variables están asociados con un aumento en el GO_P, mientras que valores bajos tienden a reducir la predicción. Además, la dispersión de los valores SHAP sugiere que algunas variables tienen un efecto no lineal en la predicción, lo que resalta la complejidad de las relaciones dentro del modelo.



Discusión

El análisis de la importancia de variables y los valores SHAP sugieren que la infraestructura (construcción residencial y no residencial) y la inversión en maquinaria y equipo de transporte son factores clave en la producción manufacturera. Esto es consistente con estudios previos que resaltan la importancia de la inversión en activos físicos en el crecimiento de la industria manufacturera.

Por otro lado, la baja influencia de la variable Ip_IT (Equipos computacionales) puede deberse a la naturaleza del índice utilizado o a que su impacto en la producción puede manifestarse en el largo plazo de manera indirecta.

En términos de desempeño del modelo, el Random Forest demostró ser la mejor opción debido a su menor error y mayor capacidad de generalización en comparación con XGBoost.

5. Conclusiones

Este estudio analizó los factores que influyen en el Valor Bruto de Producción (GO_P) en la industria manufacturera entre 1990 y 2018, utilizando modelos de aprendizaje automático para identificar los principales determinantes de la producción. A partir del análisis comparativo entre Random Forest y XGBoost, se concluyó que Random Forest ofrece un mejor desempeño predictivo, con un menor error cuadrático medio ($MSE = 7.5132$) y un coeficiente de determinación más alto ($R^2 = 0.9922$), lo que lo convierte en la mejor opción para modelar esta relación.

El análisis de importancia de variables y los valores SHAP revelaron que los factores con mayor impacto en la producción manufacturera son la inversión en equipo de transporte (Ip_TraEq), estructura residencial (Ip_RStruc), construcción no residencial (Ip_OCon), servicios laborales (LAB_QI) y otra maquinaria y equipos (Ip_OMach). Estos resultados destacan la relevancia de la infraestructura y el capital físico en el crecimiento del sector manufacturero. En contraste, la inversión en equipos computacionales (Ip_IT) tuvo la menor influencia en la predicción, lo que



sugiere que su impacto puede ser más indirecto o depender de otros factores no considerados en el modelo.

Los hallazgos de este estudio pueden ser útiles para la toma de decisiones en políticas de inversión en la industria manufacturera, priorizando aquellos activos que tienen una mayor correlación con la producción. Además, el uso de aprendizaje automático demostró ser una herramienta poderosa para analizar relaciones complejas entre múltiples variables económicas. Futuras investigaciones podrían ampliar este análisis incorporando efectos de cambios tecnológicos, políticas económicas o variaciones en la demanda global para obtener una visión más completa de la dinámica de la producción manufacturera.

6. Referencias

- Aspe Armella, P. (1993). EL CAMINO MEXICANO DE LA TRANSFORMACIÓN ECONÓMICA.
- Banco Mundial (2024). World Development Indicators. The World Bank.
- Banco Mundial (2024). Informe Económico de América Latina. The World Bank.
- Calva, J. L. (2019). La economía mexicana en su laberinto neoliberal. *El trimestre económico*, 86(343), 579-622.
- Cárdenas, E. (2010). La economía mexicana en el dilatado siglo XX, 1929-2009. *Historia económica general de México*, 503-548.
- Fajnzylber, F. (1990). Industrialización en América Latina: de la caja negra" al" casillero vacío": comparación de patrones contemporáneos de industrialización". Cepal.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
<http://www.deeplearningbook.org>
- Kang, H. S., Lee, J. Y., Choi, S., Kim, H., Park, J. H., Son, J. Y., ... & Do Noh, S. (2020). Smart manufacturing: Past research, present findings, and future directions. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 7(1), 111-128.
- LAKLEMS (2021). LAKLEMS: Crecimiento Económico y Productividad en América Latina. Julio de 2021. Base de datos disponible en: <http://www.laklems.net/>



- Lee, C., & Lim, C. (2021). From technological development to social advance: A review of Industry 4.0 through machine learning. *Technological Forecasting and Social Change*, 167, 120653.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill. ISBN 0071154671, 9780071154673
- Organización para la Cooperación y el Desarrollo Económicos (OECD). (s.f.). *OECD – Better Policies for Better Lives*. OECD. Recuperado el [12 de enero 2024], de <https://www.oecd.org/>
- Salinas, Carlos (2000). *México, un paso difícil a la modernidad*. Plaza & Janés Editores, Barcelona.
- Tello, C. (1980). *La política económica en México: 1970-1976, siglo xxi editores*.
- Villarreal, R., & Ramos, R. (2001). La apertura de México y la paradoja de la competitividad: hacia un modelo de competitividad sistémica. *Comercio Exterior*, 51(9), 772-788.
- Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research*, 4(1), 23-45.
- Zhang, C., Xu, W., Liu, J., Liu, Z., Zhou, Z., & Pham, D. T. (2020). A reconfigurable modeling approach for digital twin-based manufacturing system. *Procedia CIRP*, 93, 1679-1684.