Machine Learning para el mejoramiento de calidad en industria manufacturera *

Angie Paola Correa Sepúlveda *Universidad de Antioquia*
Olga Cecilia Usuga Manco *Universidad de Antioquia*

Resumen: El objetivo del trabajo es presentar la metodología de mejoramiento de calidad en una industria manufacturera a partir de la utilización de técnicas de Machine Learning en el lenguaje de programación R. En particular se aplicaron diversas técnicas de Machine Learning con el objetivo de predecir un indicador de producción muy importante en una compañía del sector de alimentos, aprovechando la información capturada en distintas etapas del proceso de elaboración del producto principal de dicha compañía. Las técnicas que presentaron mejor desempeño fueron modelos de regresión lineal, modelos GAMLSS y máquinas de soporte vectorial. Adicionalmente, se desarrolló una aplicación Shiny para la presentación del análisis.

Keywords: Machine Learning, Manufactura, Shiny.

Introducción

Conforme la tecnología avanza a pasos agigantados, la industria de la manufactura se enfrenta al reto de recolectar, comprender y analizar una gran cantidad de datos con el objetivo de ser más eficientes operativamente y responder rápidamente a las necesidades de los consumidores (Nhuch, 2017). Por tal razón, la analítica de datos ofrece la oportunidad de extraer información valiosa y crear modelos predictivos no sólo para analizar comportamientos históricos sino también predecir diversas variables teniendo en cuenta múltiples escenarios.

Autores como Yuan (2018) describen que la manufactura se está contagiando rápidamente del auge de la inteligencia artificial gracias a que su incorporación trae beneficios en costos de operación e incremento en la productividad. Por esa razón, las empresas se han dado a la tarea de recolectar grandes volúmenes de datos, procesarlos y encontrar patrones para detectar y predecir fallas; todo esto utilizando herramientas de deep learning y machine learning.

Particularmente, en este caso de estudio los esfuerzos se concentraron en la incorporación de la analítica de datos para la toma de decisiones usando el lenguaje de programación R R Core Team (2019) en el marco del plan estratégico de una compañía del sector de alimentos que tiene como pilar adaptar el modelo de la Industria 4.0 en aras de ser más competitivos. El trabajo presentado se divide en tres partes: metodología, resultados y conclusiones.

Metodología

Para el análisis del indicador de sobrepeso en las galletas de soda producidas por una compañía del sector de alimentos se desarrollaron las siguientes etapas: preparación de datos, análisis descriptivo de datos, modelación y visualización; bajo el modelo tradicional mostrado en la Figura 1.

* **Autor de contacto**: angiepcorreas@gmail.com

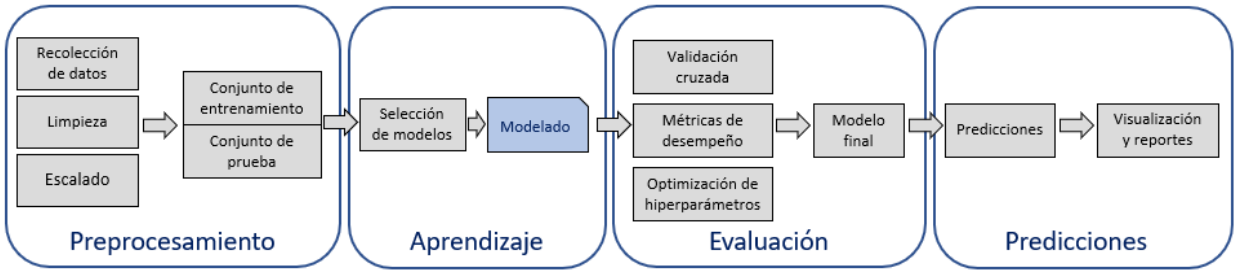

Figure 1: Metodología para la construcción de modelos.

En la preparación de los datos se hizo una exploración inicial para la verificación de valores faltantes y presencia de outliers, pero no hubo necesidad de hacer ningún tratamiento, ya que, en la extracción y carga, los datos ya estaban preprocesados. Sin embargo sí fue necesario codificar las variables categóricas y esto se hizo a través de la función `factor()`. En algunos casos fue necesario escalar las variables cuantitativas para evitar que tuvieran más peso unas sobre otras debido a su escala de medición; para ello, se hizo uso del paquete **caret** (Kuhn, 2019).

En el análisis exploratorio de datos se identificaron relaciones importantes para el indicador del Sobrepeso a partir de una matriz de dispersión y correlación. Así que para ello se usaron los paquetes **GGally** (Schloerke et al., 2018), **Scatterplot3d** (Ligges and Mächler, 2003) y **car** (Fox and Weisberg, 2019). Posterior a esto, se listaron los modelos a ajustar.

En la predicción del indicador del sobrepeso se usaron modelos de regresión, modelos GAMLSS, máquinas de soporte vectorial y bosques aleatorios. Para la predicción se usaron los paquetes **lmtest** (Zeileis and Hothorn, 2002), **caret** (Kuhn, 2019), **rpart** (Therneau and Atkinson, 2019), **randomForest** (Liaw and Wiener, 2002), **e1071** (Meyer et al., 2019), **gamlss** (Rigby and Stasinopoulos, 2005). Para analizar el desempeño de los modelos se usó el paquete **MASS** (Venables and Ripley, 2002) y **caret** (Kuhn, 2019) para la validación cruzada.

Para la visualización del análisis exploratorio y modelos predictivos se usaron los paquetes **shiny** (Chang et al., 2019), **shinydashboard** (Chang and Borges Ribeiro, 2018), **shinythemes** (Chang, 2018), **ggplot2** (Wickham, 2016), **dplyr** (Wickham et al., 2019) y **tidyr** (Wickham and Henry, 2019).

Resultados

De manera descriptiva se realizaron pruebas gráficas y analíticas para verificar la distribución del Sobrepeso, como se observa en la Figura 2. Como el objetivo es predecir el Sobrepeso de las galletas de soda (expresado en términos porcentuales), lo usual es pensar en un modelo adaptado a datos que toman valores desde cero hasta uno. Sin embargo, este indicador puede tomar valores inferiores a cero y esto sucede cuando se produce menos de lo esperado o la galleta tiene un bajo peso, además, es casi improbable que el sobrepeso tome valores muy grandes, por ejemplo, mayores al 20%. Por lo tanto, no se consideró adecuado ajustar un modelo a datos porcentuales y mejor se optó por utilizar otras técnicas dependiendo de la distribución de los datos.

Se planteó un modelo de regresión lineal múltiple considerando variables como el mes, día de

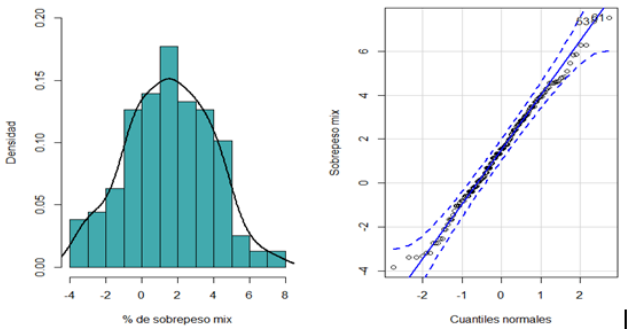

Figure 2: Distribución del porcentaje de sobrepeso.

la semana, turno de trabajo (turno 1, turno 2 y turno 3), peso de 10 galletas, resistencia promedio de la galleta, ancho, calibre, PH y humedad. Previo al ajuste del modelo se hizo una selección de variables a través de una combinación de los métodos de selección forward y backward, donde las variables significativas fueron: mes, turno, peso de 10 galletas, resistencia promedio y calibre de las galletas.

También, se decidió aplicar modelos GAMLSS para predecir el sobrepeso en función de las mismas variables independientes descritas previamente. Se ajustaron las cuatro distribuciones estadísticas que mejor explicaron el comportamiento del sobrepeso, sin incluir las variables independientes; y dichas distribuciones fueron: normal, normal generalizada o power exponential, power exponential tipo 2 y skew-normal tipo 2; en ese orden. Luego, se ajustaron los cuatro modelos de las mejores distribuciones y se compararon entre sí a partir del Criterio de Información de Akaike (AIC) y el comportamiento observado en los worm plots calculados con la función `wp()`. El mejor modelo GAMLSS fue aquel con variable respuesta normal y las variables significativas fueron iguales a las del modelo de regresión lineal, a un nivel de significacia de 0.05. Todo lo anterior se realizó con el paquete **gamlss** (Rigby and Stasinopoulos, 2005).

También se plantearon dos modelos adicionales: un modelo de bosques aleatorios y otro de máquinas de soporte vectorial. Para ello, se dividió el conjunto de datos original en dos grupos: uno de entrenamiento, que correspondió al 80% del conjunto de datos original, y otro conjunto de validación con el 20% de datos restantes. Se utilizaron 1000 árboles para el modelo de bosques aleatorios, mientras que para el modelo de máquinas de soporte vectorial se optimizó el valor del costo, cuyo mejor desempeño se da con un valor igual a 1. Tanto en el bosque aleatorio como en las máquinas de soporte vectorial la variable con mayor importancia fue el mes, seguido del calibre de la galleta, mientras que la variable con menor importancia fue el día de la semana; lo cual coincide con el modelo de regresión lineal y los modelos GAMLSS, previamente ajustados.

Con base en el error cuadrático medio y la correlación con la variable respuesta, el modelo con mejor desempeño fue la regresión lineal múltiple con una correlación de 0.75 y un error cuadrático medio de 2.44 (Ver Tabla 1).

Por último, se desarrolló una aplicación Shiny para mostrar el análisis descriptivo y las predicciones del indicador de Sobrepeso, y de esta forma, facilitar el entendimiento de los resultados del análisis y, al mismo tiempo, entregar una herramienta para la toma de decisiones en el proceso productivo.

Table 1: Métricas de evaluación para los modelos ajustados.

Modelo	Error cuadrático medio	Correlación
Regresión lineal múltiple	2.44	0.75
Bosque aleatorio	4.28	0.43
Máquinas de soporte vectorial	3.51	0.58
GAMLSS	2.48	0.74

Conclusiones

La analítica de datos se ha convertido en una de las herramientas más usadas en la industria debido a su eficiencia en el análisis profundo de información y su capacidad para acoplarse al entorno dinámico y caótico que suponen los sistemas de producción.

En este trabajo se presentaron aplicaciones de modelos predictivos enfocados a la predicción del indicador de calidad sobrepeso del producto; indicador que día a día se calcula y se presenta para conocer el estado del proceso y soportar la toma de decisiones. Este indicador es uno de los más importantes, ya que cuantifica el peso por encima o por debajo de las especificaciones, lo que incurre en costos adicionales de materia prima y material de empaque.

El modelo de regresión lineal múltiple fue el que explicó mejor el sobrepeso por encima de las máquinas de soporte vectorial, bosques aleatorios y modelos GAMLSS, siendo este último muy parecido en cuanto a resultados al modelo de regresión lineal múltiple. Las variables con mayor significancia fueron el mes, el turno y el calibre de la galleta, siendo el turno 3 el que trabaja con mayor porcentaje de sobrepeso.

Por último, con el presente trabajo se logró crear una cultura de análisis de datos que servirá para marcar un comienzo en herramientas como el machine learning enfocado hacia la manufactura, y se incentivó el uso de herramientas de análisis estadístico como R, cuyo potencial es enorme para el control de procesos y el aprendizaje de máquinas.

References

Chang, W. 2018. *shinythemes: Themes for Shiny*. R package version 1.1.2.

Chang, W. and B. Borges Ribeiro. 2018. *shinydashboard: Create Dashboards with 'Shiny'*. R package version 0.7.1.

Chang, W., J. Cheng, JJ. Allaire, Y. Xie and J. McPherson. 2019. *shiny: Web Application Framework for R*. R package version 1.3.2.

Fox, J. and S. Weisberg. 2019. *An R Companion to Applied Regression*. Third ed. Thousand Oaks CA: Sage.

Kuhn, M. et al. 2019. *caret: Classification and Regression Training*. R package version 6.0-84.

Liaw, A. and M. Wiener. 2002. "Classification and Regression by randomForest." *R News* 2(3):18–22.

Ligges, U. and M. Mächler. 2003. "Scatterplot3d - an R Package for Visualizing Multivariate Data." *Journal of Statistical Software* 8(11):1–20.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel and F. Leisch. 2019. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-2.

Nhuch, M. 2017. *Transformando con datos la Industria de Manufactura*.
URL: *https://sg.com.mx/revista/50/transformando-datos-la-industria-manufactura*

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rigby, R. and D. Stasinopoulos. 2005. "Generalized additive models for location, scale and shape,(with discussion)." *Applied Statistics* 54:507–554.

Schloerke, B., J. Crowley, D. Cook, F. Briatte, M. Marbach, E. Thoen, A. Elberg and J. Larmarange. 2018. *GGally: Extension to 'ggplot2'*. R package version 1.4.0.

Therneau, T. and B. Atkinson. 2019. *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.

Venables, W. and B. Ripley. 2002. *Modern Applied Statistics with S*. Fourth ed. New York: Springer.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. and L. Henry. 2019. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.8.3.

Wickham, H., R. François, L. Henry and K. Müller. 2019. *dplyr: A Grammar of Data Manipulation*. R package version 0.8.3.

Yuan, Y. et al. 2018. "Artificial Intelligent Diagnosis and Monitoring in Manufacturing." *arXiv preprint arXiv:1901.02057* .

Zeileis, A. and T. Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2(3):7–10.