UNIVERSIDAD DE ANTIOQUIA

MODELOS PREDICTIVOS PARA INDICADORES DE PRODUCCIÓN DE LA COMPAÑÍA DE GALLETAS NOEL

Autor
Angie Paola Correa Sepúlveda

Universidad de Antioquia
Facultad de Ingeniería, Departamento de Ingeniería Industrial
Medellín, Colombia
2019

MODELOS PREDICTIVOS PARA INDICADORES DE PRODUCCIÓN DE LA COMPAÑÍA DE GALLETAS NOEL

AUTOR:
ANGIE PAOLA CORREA SEPÚLVEDA

INFORME DE PRÁCTICA
COMO REQUISITO PARA OPTAR AL TÍTULO DE:
INGENIERA INDUSTRIAL

ASESOR INTERNO:
PhD. OLGA CECILIA ÚSUGA MANCO
PROFESORA ASOCIADA

ASESOR EXTERNO:
LUIS GUILLERMO MAYA HERNÁNDEZ
JEFE DE INFORMACIÓN DE PRODUCCIÓN – COMPAÑÍA DE GALLETAS NOEL

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA, DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
MEDELLÍN, COLOMBIA
2019

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

MODELOS PREDICTIVOS PARA INDICADORES DE PRODUCCIÓN DE LA COMPAÑÍA DE GALLETAS NOEL

RESUMEN

Los procesos de manufactura son sistemas complejos, dinámicos y expuestos a comportamientos caóticos. Es por eso que deben utilizarse todas las herramientas disponibles para realizar un seguimiento, control, prevención e intervención del proceso cuando sea necesario, de manera que se minimice el riesgo de un contratiempo que impida satisfacer la demanda o se altere la calidad del producto incurriendo en pérdidas para la compañía.

Una de las herramientas más prometedoras en cuanto al análisis de procesos y eficiencia para adelantarse a situaciones futuras es la analítica de datos, que hace parte de la nueva tendencia de la Industria 4.0 y promete responder a los desafíos de la manufactura apoyándose en grandes cantidades de datos disponibles para su posterior procesamiento y presentación de resultados. En este trabajo en particular, se aplicaron diversas metodologías de machine Learning con el objetivo de predecir algunos indicadores de producción de la Compañía de Galletas Noel aprovechando la información capturada en distintas etapas del proceso de elaboración de galletas de dicha compañía.

Previo al planteamiento de modelos de predicción se realizó un análisis descriptivo con el objetivo de explorar los datos disponibles, encontrar posibles patrones y formular los modelos a utilizar de acuerdo a los objetivos planeados y las características de los datos. Posteriormente, se realizaron predicciones sobre averías en los equipos empleando algoritmos como la regresión logística, bosques aleatorios y máquinas de soporte vectorial y se encontró que el turno de trabajo no influye significativamente en la probabilidad de ocurrencia de una avería mecánica. Por otro lado, también se realizaron predicciones sobre un indicador llamado reproceso, que es la cantidad de galleta no conforme, y pudo concluirse que variables como el horno y el tipo de reproceso son las más importantes a la hora de explicar y predecir los kilogramos de galleta no conforme en el área de multiempaque. Por último, se realizaron predicciones para el sobrepeso de la galleta Saltín Fit taco x 5 en donde el mejor modelo predictivo fue la regresión lineal múltiple por encima de los bosques aleatorios, máquinas de soporte vectorial y modelos GAMLSS, y las variables significativas fueron el mes, turno, resistencia y calibre de la galleta.

Finalmente, el análisis y las predicciones fueron presentadas a la Dirección de Producción de la Compañía de Galletas Noel de una forma clara y entendible, acompañado de una formación para la correcta utilización de los modelos.

1. 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 (Miguel 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. Sin embargo, el éxito de modelos predictivos depende de la disponibilidad de datos correctos, la estructuración adecuada del problema a modelar y la evaluación precisa de las predicciones (Big Data Republic, 2017).

En la industria manufacturera en particular, el análisis de datos permite tomar decisiones en tiempo real al realizar predicciones sobre el riesgo de fallos en los equipos y así reducir costos por mantenimiento (Metalmecánica Internacional, 2017). Yuan et al. (2018) describe 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.

Otras aplicaciones del análisis de datos también pueden remitirse a la industria galletera, en donde se busca minimizar el porcentaje de galleta no conforme y evaluar el impacto de materia prima principal, como el trigo, en la durabilidad de las galletas (Cabeza Rodríguez, 2013).

Particularmente, en la Compañía de Galletas Noel recientemente los esfuerzos se están centrando en la incorporación de la analítica de datos para la toma de decisiones en el marco de su nuevo plan estratégico que tiene como pilar adaptar el modelo de la Industria 4.0 en aras de ser más competitivos (Compañía de Galletas Noel, 2018).

Actualmente, la Compañía de Galletas Noel cuenta con la Gerencia de Operaciones y la Dirección de Producción, área internamente conocida como el PIM (Procesos de Información de Manufactura), la cual se encarga de la recolección, consolidación y análisis de toda la información relacionada con los procesos operativos de la planta de galletas. La información recolectada se relaciona con: indicadores de productividad, averías de equipos, ocupación de máquinas y empleados, árbol de

pérdidas, entre otros datos que son capturados en cada etapa del proceso de elaboración de galletas. Estos indicadores de resultados son analizados a partir de reportes estáticos que simplemente cuentan el comportamiento de la línea de producción en tiempo pasado, por lo que la toma de decisiones queda sujeta a una mayor incertidumbre.

Por lo anterior, uno de los objetivos de la Dirección de Producción de la Compañía de Galletas Noel es predecir las fallas de los equipos de acuerdo a ciertas condiciones o variables de entrada y de igual forma, también se quiere predecir cuál será la cantidad de reproceso o recorte (galleta no conforme) y el porcentaje de sobrepeso (peso de la galleta por encima del objetivo) teniendo en cuenta ciertas condiciones de operación.

Es por esto, que en el presente trabajo se propondrán modelos predictivos para la estimación de los kilogramos de reproceso, el porcentaje de sobrepeso de la galleta Saltín FIT taco x 5 y las averías en equipos, teniendo en cuenta un análisis exploratorio previo y presentando los modelos de tal forma que puedan ser fácilmente entendidos por cualquier colaborador perteneciente a la Dirección de Producción de la Compañía de Galletas Noel.

2. OBJETIVOS

2.1. Objetivo general

Desarrollar modelos predictivos para la estimación de los kilogramos de reproceso en el área de multiempaque y el porcentaje de sobrepeso de la galleta Saltín FIT taco x 5, de acuerdo a condiciones específicas de operación, así como la predicción de fallas mecánicas y eléctricas en los equipos de las diferentes etapas del proceso productivo.

2.2. Objetivos específicos

- Identificar las relaciones entre las variables asociadas al reproceso y el sobrepeso, así como los efectos en las fallas de los equipos de diferentes áreas del proceso.
- Identificar los modelos predictivos que permitan predecir de forma acertada las variables de interés.
- Predecir los kilogramos de recorte en el área de multiempaque y el porcentaje de sobrepeso de la galleta Saltín FIT taco x 5, así como la probabilidad de falla en los equipos de las diferentes áreas del proceso.

- Capacitar al personal de la Dirección de Producción para crear una cultura del análisis de datos y la adecuada utilización de los modelos de predicción estimados.

3. MARCO TEÓRICO

Una de las herramientas que responde de forma eficiente a la dinámica de los procesos de manufactura es la **analítica de datos** que, según Gartner Tech (2017), se define como la examinación autónoma o semi-autónoma de datos o contenidos utilizando técnicas y herramientas sofisticadas que van más allá del análisis tradicional, y uno de los usos de la analítica de datos que ha tomado mayor fuerza es el desarrollo de **modelos predictivos** (Lee et. al. 2013), que no es más que utilizar datos históricos y nuevos para predecir un comportamiento futuro por medio de técnicas estadísticas y de computación.

En este caso en particular, los modelos predictivos tendrán especial importancia para la predicción de **averías** en los equipos, que Compañía de Galletas Noel define como la detención de la función básica de la máquina, mayor a cinco minutos, que requiere reparación y/o cambio de piezas. Dichas averías pueden ocasionar lo que en el negocio de galletas suele llamarse **reproceso o recorte,** que es un producto no conforme que se genera en el proceso productivo y puede ser incorporado o no al proceso.

Adicional a las averías, existen otras variables que pueden afectar la calidad de las galletas como el trigo, el cual influye en el reproceso y en el **sobrepeso mix**, que es la galleta que tiene un peso mayor o menor al rango estipulado como objetivo (Compañía de Galletas Noel, 2015).

Con toda la información anterior se calcula uno de los indicadores de productividad más importantes, el **OEE**; el cual indica la efectividad de los equipos en planta, relacionando el tiempo de trabajo real sobre un tiempo esperado (LeanSis, 2018). Sin embargo, en el área del **PIM** (procesos de Información de Manufactura), cuya función es la recolección y análisis de datos, siempre se realiza un análisis estático, por lo que los modelos predictivos empezarán a ser relevantes, pero no sin antes realizar un juicioso **análisis exploratorio** de los datos, con el que se hacen investigaciones preliminares sobre los mismos (Batanero et. al. 1991), y sobretodo, al final de la modelación deben presentarse de manera adecuada y entendible los resultados a través de una **visualización de datos** y, a la vez, es importante capacitar al personal para crear una cultura de la analítica de datos.

Retomando el concepto de los modelos estadísticos predictivos, existen múltiples técnicas estadísticas de predicción que de acuerdo al objetivo concreto del análisis predictivo que se quiere realizar y a las características específicas del conjunto de datos a utilizar, se ajustan correctamente. En este caso en particular, la regresión logística, regresión lineal múltiple, modelos GAMLSS, bosques aleatorios y las máquinas de soporte vectorial son modelos que se adaptan adecuadamente a los objetivos trazados en el presente trabajo y, por supuesto, a las características y particularidades identificadas en el análisis exploratorio de los conjuntos de datos utilizados.

Por otro lado, dentro del ajuste de modelos predictivos debe tenerse en cuenta una fase de procesamiento de los datos que consiste en realizar una división del conjunto de datos original en dos, un conjunto de datos de entrenamiento y otro para la validación del modelo. El set de datos de entrenamiento es el que se utiliza para entrenar el modelo, por tanto, el modelo observa y 'aprende' de este conjunto de datos para posteriormente hacer predicciones basadas en este aprendizaje. El otro conjunto de datos es el set de datos para la validación del modelo en donde se comprueba el ajuste del modelo y su capacidad predictiva. Es de aclarar que el modelo no debe aprender del conjunto de validación. Según Harrington, (2012) en su libro "*Machine Learning in action*", se recomienda que el 70 u 80% de los datos originales correspondan al conjunto de entrenamiento, pues el conjunto debe ser lo suficientemente representativo como para que el modelo 'aprenda' completamente.

A continuación, se explicarán brevemente los conceptos de los modelos predictivos que se utilizan en el desarrollo del presente trabajo.

- **Regresión lineal múltiple:**

El modelo de regresión lineal múltiple es una extensión del modelo de regresión lineal simple en el que el valor de la variable respuesta Y se determina a partir de un conjunto de variables independientes $(X_1, X_2, \dots, X_n)$.

Los modelos de regresión lineal múltiple presentan la siguiente estructura:

$$\boldsymbol{Y}_i = (\beta_0 + \beta_1 \boldsymbol{X}_{1i} + \beta_2 \boldsymbol{X}_{2i} + \cdots + \beta_n \boldsymbol{X}_{ni}) + e_i \qquad (1)$$

donde β_0 es el intercepto; β_i es el efecto promedio que tiene el incremento en una unidad de la variable independiente $\boldsymbol{X}_i$ sobre la variable respuesta $\boldsymbol{Y}$, siempre y cuando el resto de variables independientes se mantengan constantes; y e_i es el residuo o diferencia entre el valor observado y el valor estimado por el modelo.

Para la aplicación de un modelo de regresión lineal múltiple deben cumplirse los siguientes supuestos: los residuos deben distribuirse de forma normal con media igual a cero y varianza constante (homocedasticidad), esto es, $e_i \sim N(0, \sigma^2)$. Si la varianza es constante, no debe observarse ningún patrón en la distribución de los residuos. Por otro lado, los errores deben ser independientes entre las observaciones, además no debe haber multicolinealidad, esto es, las variables predictoras deben ser independientes entre sí, pues la colinealidad ocurre cuando una variable independiente está linealmente relacionada con otra u otras variables independientes del modelo o es combinación lineal de otra variable predictora (Montgomery, Peck & Vining, 2012).

- **Regresión logística binaria:**

La regresión logística binaria es un tipo de análisis de regresión cuyos orígenes se remontan a la década de los sesenta cuando Confield, Gordon y Smith (1961) introdujeron el concepto bajo la premisa de que pudiera ser utilizada cuando se quisiera predecir una variable categórica con dos niveles o dos valores posibles, a partir de un conjunto de variables independientes que pueden ser continuas o categóricas. Esta definición es un tanto parecida a la de la regresión lineal, sin embargo, aplicar dicho modelo de regresión carecerá de sentido cuando la variable a explicar solamente pueda tomar dos valores, pues al evaluar la función para valores puntuales de las variables independientes se obtendrán valores diferentes a 0 y 1, que es como estará codificada la variable respuesta al ser de tipo binario (Hoffman, 2015).

En general, el modelo de Regresión Logística Binaria puede formularse de la siguiente manera:

$$P(Y = k \,|X = x) = \frac{e^{\beta_0 + \beta_1 X}}{1 + \ e^{\beta_0 + \beta_1 X}} \qquad (2)$$

donde $P(Y = k \,|X = x)$ puede interpretarse como la probabilidad de que la variable categórica Y adquiera el valor k (nivel de referencia usualmente codificado como 1), dado que la variable predictora X adquiere el valor x.

De manera más simple y familiarizada con el modelo de regresión lineal comúnmente conocido, la función anterior puede ajustarse a una versión logarítmica que se conoce como el logaritmo de la razón de probabilidad, así:

$$Log\left(\frac{p(Y=k\,|X=x)}{1-p(Y=k\,|X=x)}\right) = \beta_0 + \beta_1 X \quad (3)$$

en donde la interpretación del coeficiente β_1 será similar a la de los coeficientes de la regresión lineal. Si β_1 es positivo significará que incrementos en la variable X harán que el logaritmo de la razón de probabilidades también se incremente, mientras que si el signo del coeficiente es negativo entonces el logaritmo de la razón de probabilidades disminuirá por cada unidad que se incremente la variable X.

La curva resultante de un modelo de regresión logística binaria se asemejará a la ilustrada en la Figura 1, en donde se observan los posibles valores de la variable independiente (duración de la avería) y la respectiva probabilidad de ocurrencia del evento de referencia (en este caso, una avería mecánica). Nótese que la curva abarca únicamente valores para el eje y comprendidos entre 0 y 1, pues es precisamente una de las premisas de los modelos de regresión logística que se obtiene mediante las transformaciones logarítmicas y el concepto de odds ratio (razón de probabilidades).

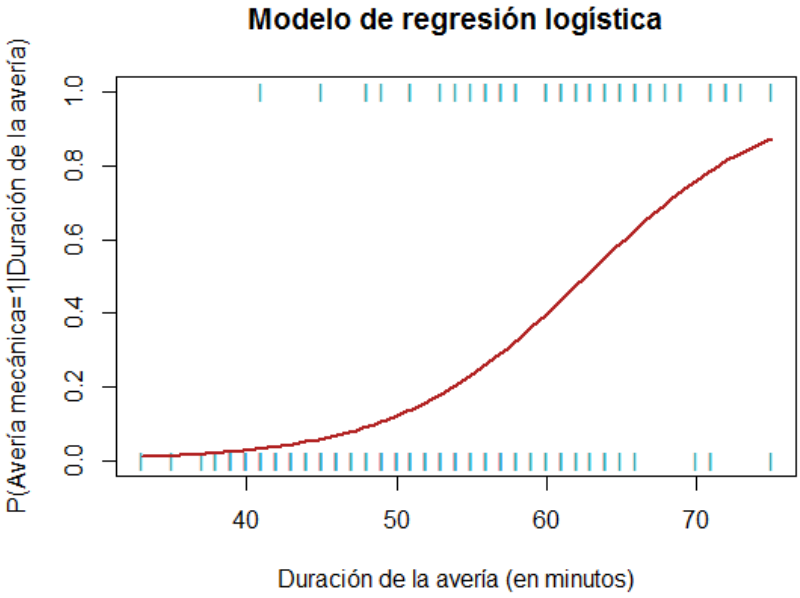

Figura 1. Curva de un modelo de regresión logística ajustado. Fuente: elaboración propia.

- **Máquinas de Soporte Vectorial:**

Dentro del ámbito del aprendizaje estadístico y Machine Learning, las Máquinas de Soporte Vectorial (SVM) se han convertido en un referente para resolver problemas de clasificación y de regresión. Si bien este método, desarrollado en los años 90, inicialmente se ideó para resolver problemas de

clasificación, su aplicación se extendió exitosamente hacia la regresión (Scholkopf, Burges & Smola,1999).

Las Máquinas de Soporte Vectorial pertenecen a la familia de clasificadores lineales, mediante una función matemática denominada *Kernel*. Por definición, una máquina de soporte vectorial construye un hiperplano o un conjunto de hiperplanos en un subespacio de $p-1$ dimensiones para separar de forma óptima los puntos de una clase de otra (Cristianini & Shawe-Taylor, 2000). Por ejemplo, en la Figura 2 se muestra el hiperplano de un espacio bidimensional, por lo que la ecuación que describe dicho hiperplano es una recta, en este caso $3x_2 + 2x_1 + 1 = 0$, en donde la región azul representa el espacio en el que se encuentran todos los puntos para los que $3x_2 + 2x_1 + 1 > 0$, mientras que la región roja representa el espacio de los puntos para los que $3x_2 + 2x_1 + 1 < 0$. De esta forma, aquellos puntos que estén situados en la región azul pertenecerán a la categoría de dicha región, mientras que las observaciones o puntos localizados en la región roja pertenecerán a la otra categoría, en caso de que se trate de un problema de clasificación binaria. Si se tratara de un problema de regresión, se utilizarían vectores de soporte para regresión, sin embargo, el concepto no cambia y se haría uso de las funciones de Kernel.

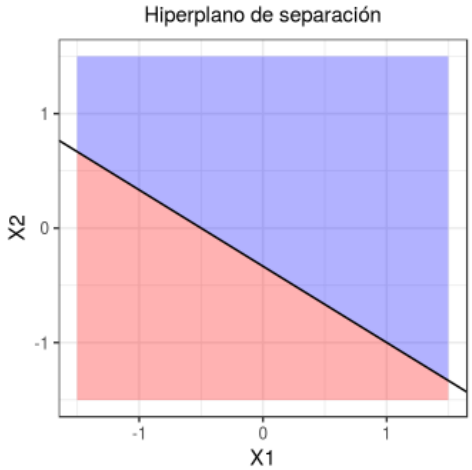

Figura 2. Hiperplano de separación bidimensional. Tomada de https://rpubs.com/Joaquin_AR/267926 .

- **Bosques aleatorios:**

Es una combinación de árboles predictores, pues es un modelo de clasificación o regresión (dependiendo del problema) que funciona creando múltiples árboles durante la etapa de aprendizaje para así conseguir una mejor predicción que la que se conseguiría con un solo árbol.

Para los problemas de clasificación, el modelo *Random Forest* o bosque aleatorio se basa en un conjunto de árboles de clasificación, en donde una

muestra de los datos originales entra al árbol y es sometida a una serie de test binarios, o sea de tipo Si/No, en cada nodo hasta llegar a una hoja en la que se encuentra la respuesta, a ese nodo terminal se le asigna una etiqueta. Este proceso es repetitivo en todos los árboles, y la etiqueta que tenga la mayor cantidad de incidencias será la predicción final (Kotu & Deshpande, 2019). De manera ilustrativa, el funcionamiento de los bosques aleatorios se muestra en la Figura 3, en donde se observa que la mayoría de los árboles clasifican un registro específico en una clase o categoría denominada "B", por tanto, la predicción o clasificación final para esa observación será en efecto dicha categoría.

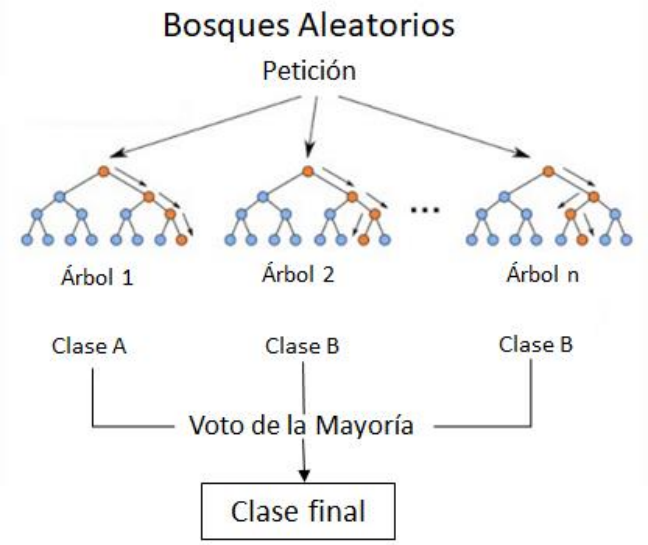

Figura 3. Esquema del funcionamiento del modelo de bosques aleatorios. Tomado de https://medium.com/datos-y-ciencia/aprendizaje-supervisado-introducci%C3%B3n-a-la-clasificaci%C3%B3n-y-principales-algoritmos-dadee99c9407

En caso de enfrentarse a un problema de regresión, la metodología no es muy distinta. En un bosque aleatorio de regresión se seleccionan al azar las observaciones y características (variables independientes) para construir varios árboles de regresión y luego promediar los resultados, así que mientras en un problema de clasificación cada árbol arroja una clase y la clase final es definida por la clase con mayor frecuencia arrojada por lo árboles, en un bosque aleatorio de regresión cada árbol arrojará una predicción numérica que es calculada a partir de una serie de test binarios teniendo en cuenta las variables independientes seleccionadas al azar al interior del árbol. La predicción final será entonces el promedio de las predicciones numéricas arrojadas por cada árbol.

- **Modelos Aditivos Generalizados de Localización, Escala y Forma:**

Los modelos aditivos generalizados para localización, forma y escala (GAMLSS) son modelos de regresión semi-paramétricos que fueron introducidos por Rigby y Stasinopoulos (2005) como una forma superar las limitaciones asociadas a los Modelos Lineales Generalizados (GLM) y los Modelos Aditivos Generalizados (GAM). Los modelos GAMLSS tienen la facilidad de que la distribución de la variable respuesta no tiene que pertenecer a la familia exponencial y puede ser altamente sesgada, y además permiten modelar todos los parámetros de la variable de interés en función de variables independientes.

Los modelos GAMLSS asumen que las observaciones son independientes. Los parámetros μ_i y σ_i corresponden a los parámetros de localización y escala, mientras que ν_i y τ_i son parámetros de forma (Stasinopoulos & Rigby, 2007).

La estructura general de los modelos GAMLSS se expresa a continuación:

$$g_1(\boldsymbol{\mu}) = \boldsymbol{\eta}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \boldsymbol{Z}_{j_1}\boldsymbol{\gamma}_{j_1} \qquad (4)$$

$$g_2(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \boldsymbol{Z}_{j_2}\boldsymbol{\gamma}_{j_2} \qquad (5)$$

$$g_3(\boldsymbol{\nu}) = \boldsymbol{\eta}_3 = \boldsymbol{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \boldsymbol{Z}_{j_3}\boldsymbol{\gamma}_{j_3} \qquad (6)$$

$$g_4(\boldsymbol{\tau}) = \boldsymbol{\eta}_4 = \boldsymbol{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \boldsymbol{Z}_{j_4}\boldsymbol{\gamma}_{j_4} \qquad (7)$$

donde $g_k(\cdot)$ es una función de enlace conocida para $k = 1, \dots, 4$; $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, $\boldsymbol{\nu}$ y $\boldsymbol{\tau}$ son vectores n-dimensionales; $\boldsymbol{X}_k$ son matrices de orden $n \times J'_k$ asociadas a los efectos fijos β_k; mientras que $\boldsymbol{Z}_{jk}$ son matrices de orden $n \times q_{jk}$ asociadas a los efectos aleatorios γ_{jk}; J'_k es el número de variables independientes utilizadas en la parte fija del predictor lineal $\boldsymbol{\eta}_k$ y J_k representa el número de efectos aleatorios en $\boldsymbol{\eta}_k$ (Hernández, Naranjo & Monsalve, 2017).

Una vez ajustados los parámetros correspondientes de la distribución especificada, puede hacerse una interpretación para la media y varianza de la variable respuesta de acuerdo a la distribución y su función de enlace. Por ejemplo, el valor esperado para la distribución normal está dado por $E(Y) = \mu$, usando como función de enlace la función identidad, mientras que la varianza está dada por $Var(Y) = \sigma^2$, utilizando log como función de enlace.

Para comparar distintos modelos GAMLSS según su ajuste se utilizó el Worm plot que es una herramienta de diagnóstico para visualizar qué tan bien se un modelo estadístico se ajusta a los datos. El Worm plot es una modificación del gráfico qqplot en donde los valores del eje vertical corresponden a la diferencia entre la coordenada y la coordenada x del conocido qqplot, lo que hace que se cree un gráfico sin tendencia estocástica con una secuencia de puntos que, entre más plana sea, mejor asustado será el modelo (Buuren & Fredriks, 2001).

- **Evaluación de modelos predictivos**

Los modelos anteriores fueron aplicados por separado y fueron evaluados para analizar su eficiencia y capacidad predictiva a partir de algunas pruebas dependiendo del modelo ajustado.

Para evaluar la eficiencia del modelo de regresión logística binario, se utilizó la curva ROC, el test de razón de verosimilitud y la matriz de confusión. La curva ROC es una representación gráfica de la sensibilidad frente la especificidad para un sistema clasificador binario, siendo la especificidad y sensibilidad calculados a partir de la tasa de verdaderos positivos (éxitos), verdaderos negativos (rechazos correctos), falsos positivos (error tipo I) y falsos negativos (error tipo II) (Prati et al., 2008), es decir, se evalúa la capacidad del modelo para diferenciar entre las dos clases o niveles de la variable respuesta. La curva ROC también permite establecer un punto de corte para clasificar las observaciones; a valores superiores a dicho punto de corte las observaciones serán clasificadas en una categoría, mientras que a valores inferiores el modelo determinará que las observaciones pertenecen a la otra categoría (Franco & Molina, 2007).

Por su parte, el test de *Likelihood ratio* o *razón de verosimilitud* se utilizó para comprobar la significancia del modelo ajustado en general. Esta prueba compara el ajuste del modelo completo (con todas las variables predictoras) con el ajuste de un modelo 'nulo', es decir, sin predictores (Bewick, Cheek & Ball, 2005), por tanto, este test es análogo a la prueba F de Fisher del modelo de regresión lineal. Si el valor-p arrojado por la prueba es inferior al nivel de significancia dado, entonces hay evidencia de que al menos una de las variables independientes contribuye o explica la variable respuesta.

También, es muy común utilizar la matriz de confusión para medir la eficiencia del modelo de regresión logística binario (Ariza, Rodríguez & Alba, 2018). Una matriz de confusión es una matriz de orden $n\,x\,n$ en la que las filas corresponden a las clases reales de una variable categórica, mientras que las columnas corresponden a las clases previstas por el modelo ajustado; de esta forma, los valores de situados en la diagonal de la matriz corresponderán al número de aciertos o clasificaciones correctas que el modelo ajustado logró para cada categoría.

Por otro lado, los modelos de bosques aleatorios, máquinas de soporte vectorial y GAMLSS fueron evaluados a través del error cuadrático medio y la correlación entre las predicciones y el valor real de la variable respuesta; recordando que el error cuadrático medio es el promedio del cuadrado de la diferencia entre el valor predicho y el valor real de la variable respuesta, mientras que la correlación indica el grado de relación entre las predicciones de la variable respuesta y los valores reales de la misma (Taylor, 1990), así que un buen ajuste del modelo indicaría un coeficiente de correlación alto entre lo real y lo predicho.

Para el caso de los modelos GAMLSS y la regresión lineal múltiple se utilizó el criterio de Información de Akaike (AIC) para realizar una selección de variables y de esta forma obtener un modelo parsimonioso, es decir, más simple. Este criterio tiene como objetivo encontrar un modelo que haga mejores predicciones y para ello debe penalizar el doble uso de los datos (en la modelación y las predicciones). El AIC mide entonces la bondad del ajuste a partir de la máxima verosimilitud del modelo, y la complejidad a partir del número de parámetros (Akaike, 1974).

4. METODOLOGÍA

4.1. ETAPA I: Recolección de datos y análisis exploratorio

Un paso importante previo al inicio de la modelación es la contextualización e identificación del proceso a modelar, así que se organizaron varias reuniones para establecer los objetivos específicos del proyecto, definir el alcance y se realizaron visitas periódicas a la planta con el objetivo de conocer las etapas del proceso de elaboración de galletas e identificar en dónde se realizan mediciones y se captura la información.

Posteriormente, se hizo una identificación y recopilación de los datos disponibles, que actualmente se encuentran almacenados en el Sistema de Información de Producción de la compañía. Se consolidaron tres bases de datos: una base de datos con información sobre las averías, otra con información sobre el recorte o reproceso de galleta y una última base de datos que relacionaba información sobre el sobrepeso mix de la galleta Saltín Fit taco x 5 del horno 12, que es una de las referencias que más produce la compañía.

Para la base de datos de las averías se extrajeron datos almacenados en el Sistema de Información de Producción desde enero de 2014 hasta marzo de 2019, para un total de 6668 registros de averías en el proceso productivo de

Noel. Esta base de datos consolidada tenía variables como el mes de la avería, el día de la semana, la hora de inicio y fin de la avería, el horno en el cual se presentó la avería, el turno en operación, el tipo de avería, entre otros. Al ser demasiada información para modelar, se hizo una depuración y mediante diferentes gráficos exploratorios se priorizaron algunos tipos de averías y áreas del proceso.

En cuanto a la base de datos del reproceso de la galleta Saltín Fit taco x 5, se consolidó información de todo el recorte registrado en el año 2018 con variables como los kg de reproceso reportados, mes, día de la semana, horno, turno, tipo de reproceso (barredura, inconforme horno, recorte simple, con papel, etc.) y área donde se presentó el reproceso. También se realizó una depuración de la base de datos original teniendo en cuenta un análisis exploratorio preliminar.

Para la predicción del porcentaje de sobrepeso se extrajeron datos desde febrero hasta diciembre de 2018 de variables como el PH de la masa, la humedad, el peso de diez galletas recién salidas del horno, el ancho de la galleta, el calibre, la resistencia promedio de la galleta, el porcentaje de sobrepeso mix y también se tuvieron en cuenta variables como el mes, el día de la semana y el turno; todo lo anterior solamente para un producto (material) específico, la galleta Saltín FIT taco x 5.

La realización del análisis exploratorio en cada uno de los tres frentes definidos (averías, reproceso y sobrepeso mix) permitió identificar plenamente cada una de las variables de interés y detectar patrones de comportamiento. Se utilizaron métodos gráficos y medidas de resumen con el fin de realizar un análisis descriptivo y posteriormente determinar el modelo apropiado a ajustar.

4.2. ETAPA II: Modelación predictiva

A partir del análisis descriptivo realizado y una revisión exhaustiva de literatura de diversos modelos predictivos aplicados en la manufactura, se determinó el modelo predictivo que mejor se ajustaba a los datos disponibles y a los objetivos deseados. Por lo que en este punto se tuvieron en cuenta diferentes técnicas de Machine Learning y luego se ajustaron aquellos modelos que fueron más acordes al comportamiento de las variables dependientes.

Para la predicción de averías se utilizó un modelo de regresión logística binaria dado que la variable respuesta, en este caso el tipo de avería, es categórica con dos niveles: avería mecánica y avería eléctrica/electrónica.

Además, este modelo estimaría la probabilidad de que ocurriera una avería de tipo mecánico de acuerdo a ciertas condiciones que se presentaran en el proceso. El modelo logístico binario fue ajustado utilizando la función `glm` del paquete `stats` de R (R Core Team, 2018).

En cuanto al modelo del recorte/reproceso, se probaron varias técnicas debido a que la variable respuesta presentaba mucha variabilidad por lo que en un principio varios modelos no predecían correctamente los kg de reproceso reales. Se aplicaron técnicas como árboles de regresión, bosques aleatorios y máquinas de soporte vectorial, utilizando librerías de R como `rpart` (Therneau & Atkinson, 2018), `randomForest` (Liaw & Wiener, 2002) y `e1071` (David Meyer et al., 2017).

Por último, para la predicción del porcentaje de sobrepeso mix se utilizaron técnicas como la regresión lineal, bosques aleatorios y máquinas de soporte vectorial. Todo lo anterior, se desarrolló en el lenguaje de programación R (R Core Team, 2018).

4.3. ETAPA III: Evaluación de modelos predictivos y predicciones

Las decisiones que se tomarán a partir de los resultados de un modelo predictivo dependen de la precisión de los datos, por eso es importante usar información válida y evaluar que los modelos y predicciones realizadas sean correctos y coherentes. Por lo anterior, los modelos ajustados fueron evaluados para verificar la calidad del ajuste y su capacidad de predecir correctamente en concordancia con el día a día del proceso productivo. Para ello, se utilizaron diferentes técnicas, de acuerdo al tipo de modelo ajustado, y en ocasiones se realizaron correcciones de manera que las predicciones fueran lo más acertadas posible.

Para el modelo de regresión logística se realizó el *test Likelihood ratio* (razón de verosimilitud) que realiza una comparación entre el modelo ajustado y un modelo nulo (sin predictores), se calculó la eficiencia del modelo mediante una clasificación y una matriz de confusión, y además se trazó una curva *ROC* para verificar la proporción bajo la curva que puede predecir correctamente el modelo.

Para el resto de modelos aplicados se utilizaron dos conjuntos de datos: uno de entrenamiento, que correspondió al 80% de los datos iniciales, y otro de prueba, con el 20% restante de los datos. Para la partición de los datos en los dos conjuntos señalados se hizo uso del paquete `caret` de R (Kuhn et. al., 2019), de manera que todos los niveles de las variables categóricas

estuvieran presentes en ambos conjuntos. El conjunto de entrenamiento se utilizó para entrenar el modelo y de esta forma el algoritmo "aprendió" y detectó patrones en los datos para luego predecir la variable respuesta a partir de nuevos datos de entrada. Por su parte, el conjunto de datos de prueba se utilizó para testear el modelo una vez concluida la etapa de aprendizaje y así verificar la calidad de las predicciones. De manera común para los algoritmos de regresión, también se utilizó como medida el error cuadrático medio y la correlación.

4.4. ETAPA IV: Capacitación del personal

Las salidas de los modelos en el software por sí solas pueden llegar a ser difíciles de entender para muchas personas que no estén familiarizadas con el mismo o no tengan conocimientos previos en el campo de la Estadística, por eso, fue fundamental capacitar al personal de la Dirección de Producción de manera que los modelos de predicción fueran utilizados correctamente y se sacara el mayor provecho.

5. RESULTADOS Y ANÁLISIS

Para entender el proceso de elaboración de galletas, definir los objetivos y establecer el alcance del proyecto fue necesario organizar reuniones con el equipo del área de Información de Producción y realizar visitas periódicas a la planta con expertos del proceso. En la Figura 4 se muestra cada una de las etapas del proceso de elaboración de galletas y en muchas de esas etapas se captura información para la creación y reporte de distintos indicadores de resultado, como el reproceso o recorte que la compañía clasifica en diferentes tipos:

- Recorte con papel: aquella galleta no conforme que debe ser retirada del empaque, por lo tanto, no sólo habría desperdicio de galleta sino también de material de empaque.
- Barredura: es la galleta que cae al piso.
- Simple empaque: galleta que sale del horno a empaque y se detecta su no conformidad en el diverter (distribuidor o desviador de galleta hacia las líneas de empaque).
- No conforme horno: galleta no conforme debido a condiciones de horneo.
- Dulce cremada, chicharrón, orillo, cremada selección y cobertura chocolate: galleta no conforme debido problemas de crema o

cobertura. Este tipo de recorte no aplica a todas las referencias de galleta, sólo a aquellas que tienen crema o alguna cobertura.

El reproceso comienza a medirse a partir de la salida del horno y hasta la operación de multiempaque. Por otro lado, como se indicó previamente, el sobrepeso mix es otro de los indicadores importantes que Noel monitorea y se calcula a partir de la expresión 8 y la información necesaria para calcularlo es capturada en las operaciones de empaque individual y multiempaque.

$$Sobrepeso\ mix = \frac{kg\ producidos - kg\ esperados}{kg\ esperados} \times 100 \qquad (8)$$

Variables asociadas a la pasta o mezcla de la galleta como el PH y la humedad se capturan en la operación de mezcla y empaste, mientras que en la salida del horno de miden variables como el calibre, ancho de la galleta, peso de 10 galletas, resistencia, entre otras.

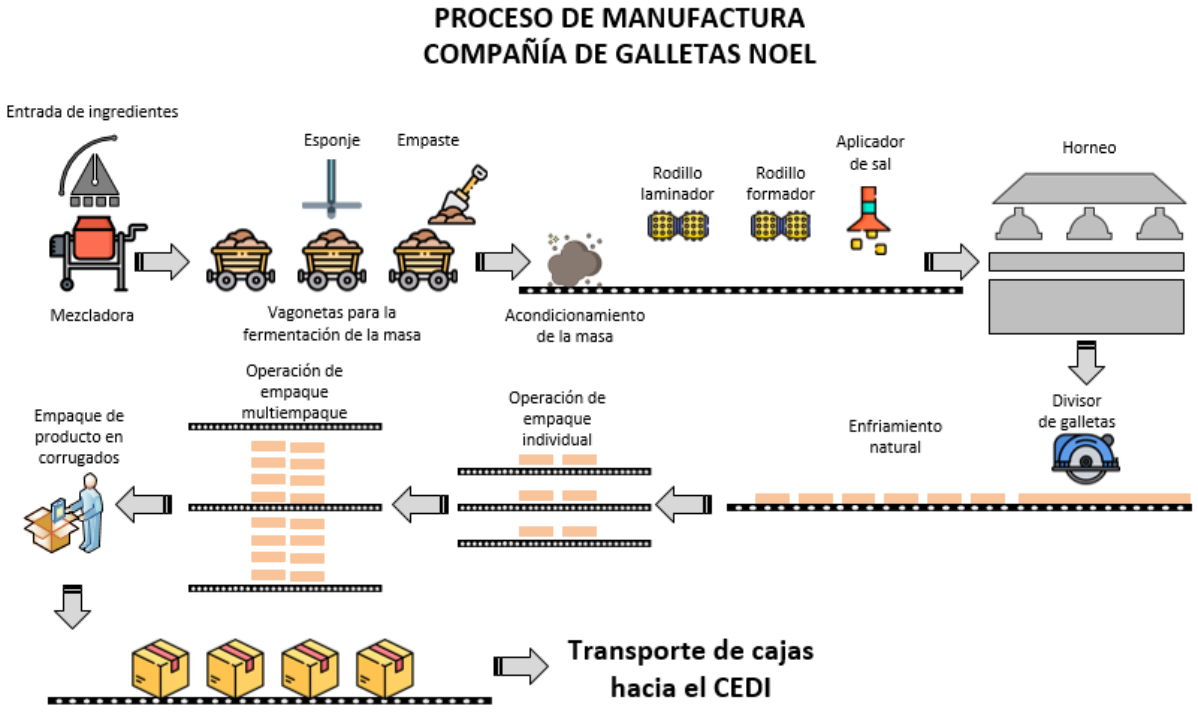

Figura 4. Proceso de producción de galletas de sal. Fuente: Compañía de Galletas Noel.

5.1. Averías en los equipos de producción.

Actualmente, en producción se manejan tres turnos, cada uno de 8 horas. El turno 1 comienza a las 5:40 a.m. y termina a la 1:40 p.m., mientras que el turno 2 va de 1:40 p.m. hasta las 9:40 p.m. y el turno 3 inicia a las 9:40 p.m. y finaliza a las 5:40 a.m. Lo normal es que se presente mayor variabilidad en todos los indicadores calculados cuando hay cambio de turno. Por ejemplo, en la Figura 5 puede constatarse que se presenta un mayor número de averías después de que se inicia un turno; esto es debido a los ajustes en máquinas y equipos que los operarios realizan según sus preferencias. Por otro lado, no se observan diferencias significativas en la frecuencia de averías entre los tres turnos (Véase la Figura 5 y Figura 6).

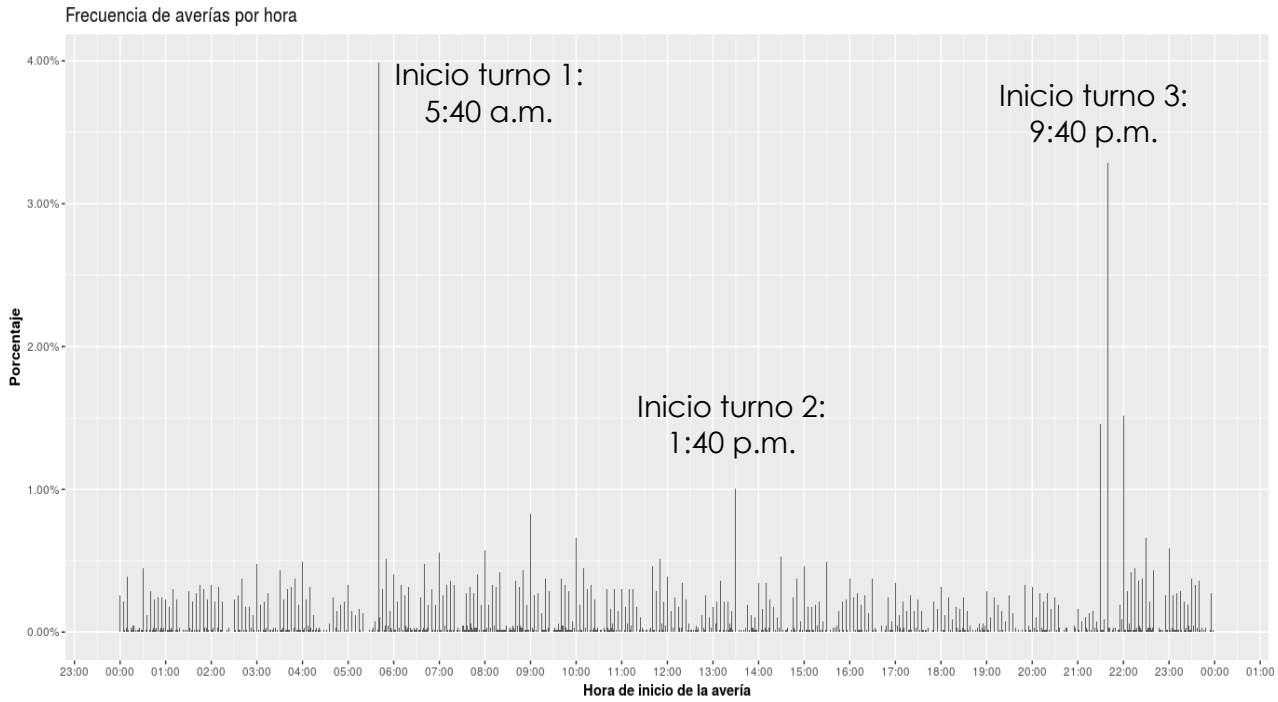

Figura 5. Frecuencia de averías por hora. Fuente: elaboración propia.

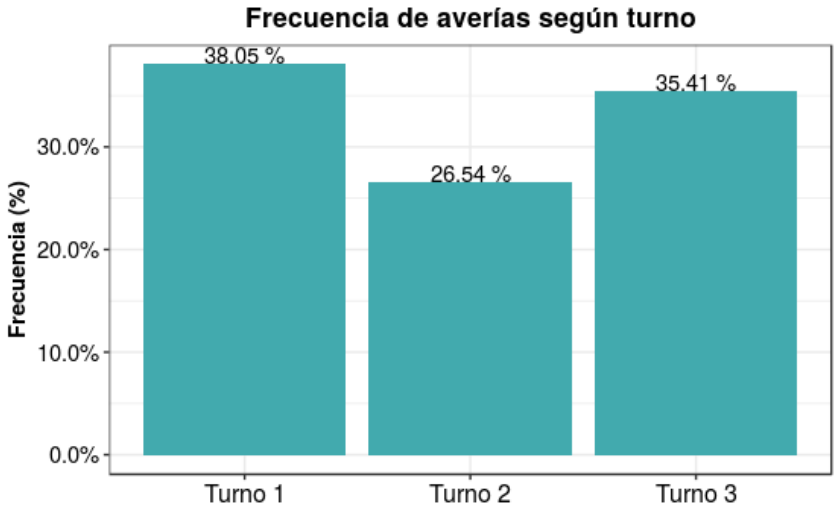

Figura 6. Frecuencia de averías por turno. Fuente: elaboración propia.

Otro aspecto que se observó en el análisis exploratorio fue la ocurrencia de averías según el día de la semana y, como se muestra en la Figura 7, conforme avanza la semana disminuyen las averías, aunque en menores proporciones, con excepciones de los días sábado y domingo en donde el porcentaje de averías es relativamente bajo, 9.73% y 2.61%, respectivamente; principalmente debido a que en los fines de semana la utilización de las máquinas es mucho menor que en el resto de días.

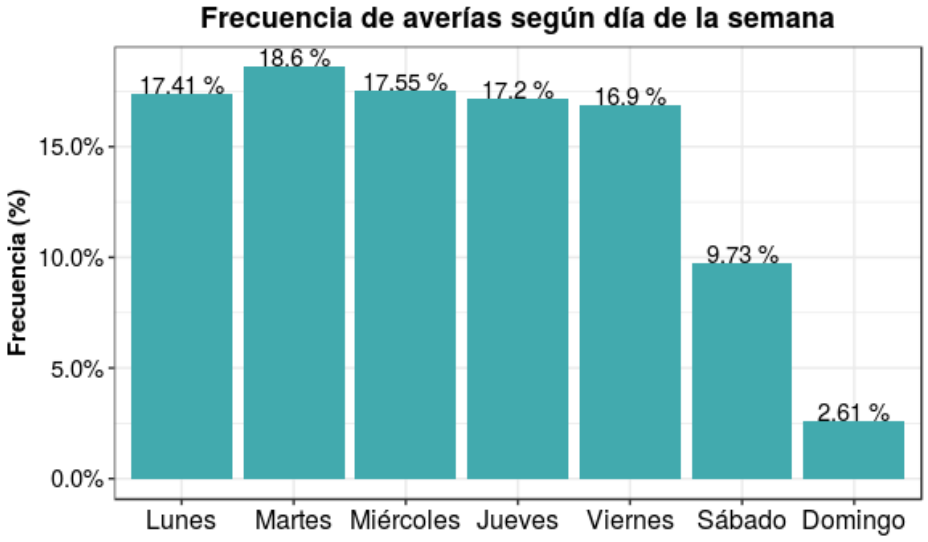

Figura 7. Frecuencia de averías según el día de la semana. Fuente: elaboración propia.

Por otro lado, actualmente se discriminan nueve tipos de averías en las máquinas de producción: mecánicas, eléctricas y electrónicas, neumáticas e hidráulicas, de instrumentación, en lonas y bandas transportadoras, en el sistema de refrigeración, en las máquinas de sticker, en marcadoras, y en servicio de aire en calderas; y en promedio aquellas que tardan más tiempo en reparar son las eléctricas/electrónicas cuya duración es de 102.9 minutos, pero las averías mecánicas son las de mayor ocurrencia con una participación del 57.29% sobre el total de averías registradas con una duración promedio de 84.4 minutos, aunque con varios registros significativamente altos alcanzando los 480 y 640 minutos de duración, como se observa en la Figura 8.

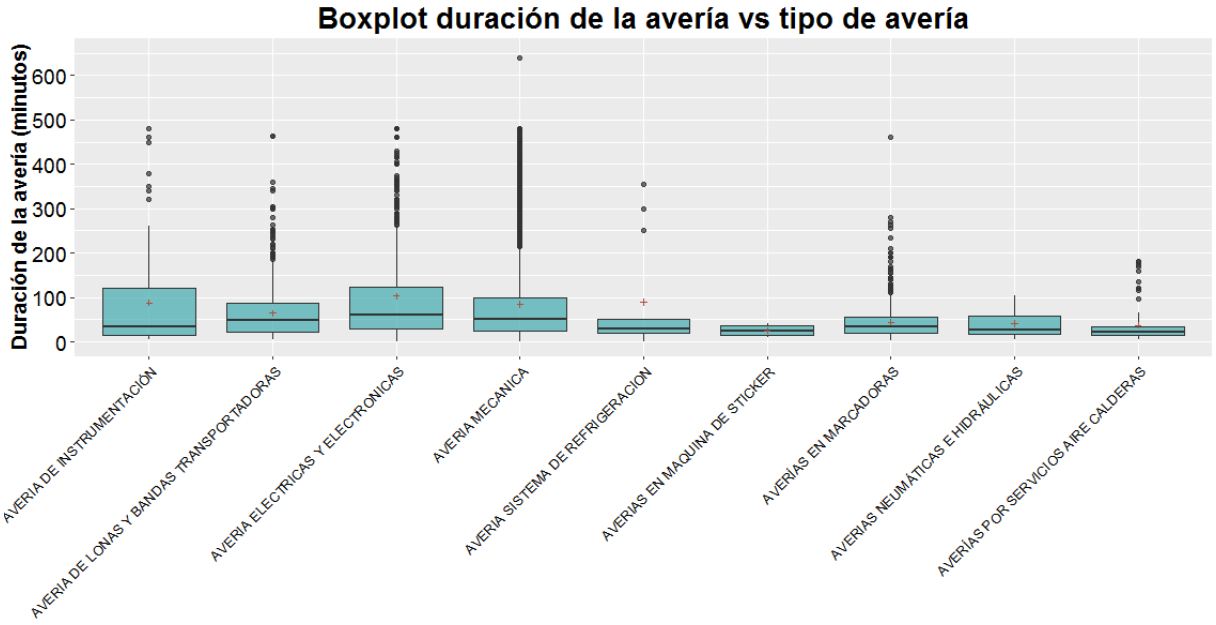

Figura 8. Boxplot para la duración de la avería según el tipo de avería. Fuente: elaboración propia.

Como existen tipos de averías con muy baja frecuencia de ocurrencia, tal es el caso de las averías neumáticas e hidráulicas, averías en máquinas de sticker y averías en el sistema de refrigeración, cuyas frecuencias relativas son de 0.09%, 0.09% y 0.19%, respectivamente, lo mejor es analizar qué tipo de averías son las más representativas para modelar, además porque de esta manera el modelo queda mejor balanceado y se reduce el riesgo de un sobreajuste.

En la Figura 9, se observa que las averías de tipo mecánico y las eléctricas/electrónicas representan el 73.9% de todas las averías registradas, y aunque también podrían incluirse las averías en lonas y bandas transportadoras, según personas inmersas en el proceso, muchas veces los colaboradores encargados de registrar en el sistema las averías ocurridas, confunden este tipo con las mecánicas e incluso piensan que algunas de las averías eléctricas corresponden a las marcadoras. Por tal razón, es conveniente tomar sólo dos niveles para el tipo de avería: las de tipo mecánico y las averías eléctricas/electrónicas.

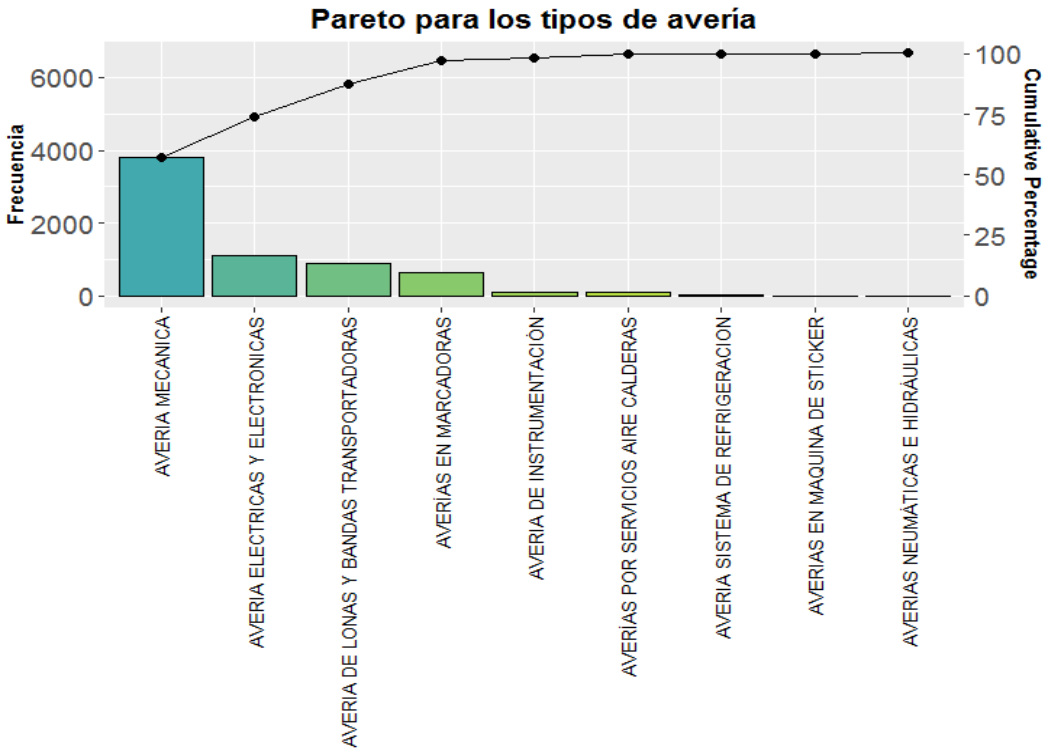

Figura 9. Pareto para los tipos de avería. Fuente: elaboración propia.

De la misma forma, existen 33 áreas dentro del proceso en las cuales se presetan averías, pero muchas de ellas tienen un porcentaje muy bajo de ocurrencia, por lo que solamente se tuvieron en cuenta las averías correspondientes a las máquinas de empaque individual, multiempaque, horneo y enfriamiento, rodillo de formación laminador, rotativa y encartonadora, quienes representan un 83.3% del total de averías (Veáse la Figura 10).

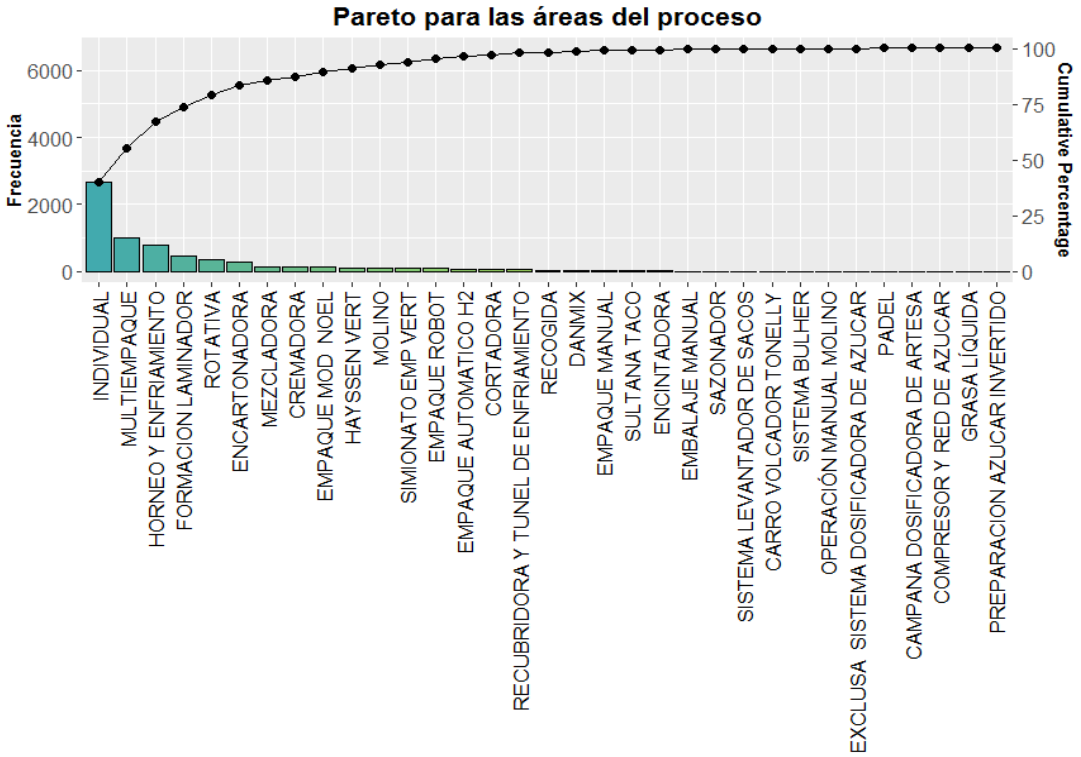

Figura 10. Pareto de la frecuencia de averías según el área del proceso. Fuente: elaboración propia.

En cuanto a hornos, en la Figura 11 se observa una mayor frecuencia de averías en los hornos Z04, Z02 y Z07, mientras que la menor ocurrencia de averías se concentra en los recursos compartidos del horno Z06 dado que este horno no trabaja al mismo ritmo que los demás.

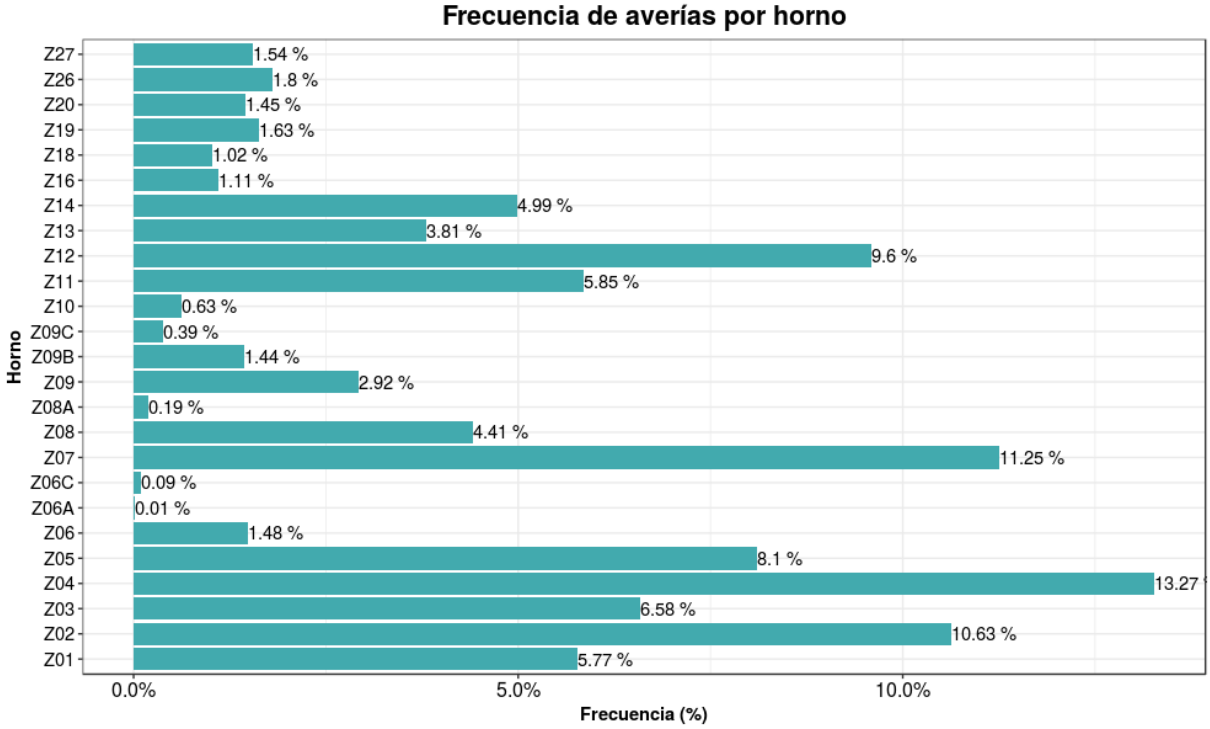

Figura 11. Frecuencia de averías por horno. Fuente: elaboración propia.

5.1.1. Regresión logística binaria para la predicción de averías

Dado que la variable respuesta, el tipo de avería, es categórica y además se restringió para dos niveles (avería mecánica y avería eléctrica/electrónica) con el objetivo de balancear mejor los datos y darle prioridad a aquellas averías que estaban ocurriendo con mayor frecuencia, se utilizó el modelo de regresión logística binaria para identificar los efectos de variables como el mes, el día de la semana, el turno, el área del proceso y el horno, y además predecir la probabilidad de que ocurra una avería mecánica de acuerdo a ciertas condiciones específicas de operación.

El modelo de regresión logística binario se especifica a continuación:

$$Log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Mes_i + \beta_2 Día_i + \beta_3 Horno_i + \beta_4 Turno_i + \beta_1 Área_i \qquad (9)$$

con $i = 6678$, donde Mes_i es el mes del año (1, …, 12), $Día_i$ es el día de la semana (lunes, …, domingo), $Horno_i$ es el horno (Z01, …, Z27), $Turno_i$ es el turno de trabajo (turno 1, turno 2 y turno 3) y $Área_i$ es el área del proceso (formación laminador, horneo y enfriamiento, empaque individual, multiempaque, rotativa y encartonadora).

A pesar de haber hecho una limpieza y un análisis preliminar, el primer modelo logístico ajustado mostró un sobreajuste debido a que el conjunto de datos utilizado para la modelación todavía presentaba un desbalance, esto es, que un 72% de los datos correspondían solamente a averías del tipo mecánico. Por esa razón, se tomó una muestra de averías mecánicas proporcional al número de datos con averías eléctricas/electrónicas.

El modelo fue nuevamente ajustado y se encontró que todas las variables fueron significativas a un nivel de significancia de 5% con excepción de la variable turno, sin embargo, en este caso se tendrá en cuenta dado que a nivel proceso es importante identificar y controlar las diferencias entre los diferentes turnos de trabajo. En la Tabla 1 se muestran todas las variables que se tuvieron en cuenta para la modelación de las averías y su valor-p correspondiente de acuerdo al test chi-cuadrado.

Tabla 1. Valor-p para las variables del modelo de averías.

Variable	Valor-p
Mes	0.005
Día	0.006
Horno	5.61e-05
Turno	0.783
Área del proceso	0.018

Los coeficientes ajustados se muestran en la Tabla 2, recordando que el modelo toma como referencia el primer nivel de cada variable categórica. Según el modelo ajustado, existe una relación significativamente positiva entre la ocurrencia de una avería mecánica y el horno Z16, es decir, que la probabilidad de que ocurra una avería mecánica es mayor en el horno Z16 que en el horno 1 (nivel de referencia), y mayor en el mes de julio comparado con el mes de enero.

Tabla 2. Coeficientes estimados del modelo de averías.

Variable	Coeficiente estimado	Variable	Coeficiente estimado
Intercepto	0.875	HornoZ04	-0.346
Mes2	-0.107	HornoZ05	-0.400
Mes3	-0.482	HornoZ06	-0.001
Mes4	-0.077	HornoZ07	0.259
Mes5	0.092	HornoZ08	-0.111
Mes6	-0.162	HornoZ09	-0.321
Mes7	0.171	HornoZ10	-1.035
Mes8	0.011	HornoZ11	-0.442
Mes9	0.125	HornoZ12	-0.307
Mes10	0.111	HornoZ13	-0.760
Mes11	0.162	HornoZ14	0.540
Mes12	-0.472	HornoZ16	13.02
DíaMartes	0.305	HornoZ27	0.132
DíaMiércoles	0.434	Turno2	-0.028
DíaJueves	0.008	Turno3	-0.082
DíaViernes	-0.091	Área Formación Laminador	-0.355
DíaSábado	0.158	Área Horneo y enfriamiento	-0.841
DíaDomingo	-0.041	Área Empaque individual	-0.712
HornoZ02	-0.061	Área Multiempaque	-0.615
HornoZ03	0.048	Área Rotativa	-1.041

5.1.2. Evaluación del modelo y predicciones

Para evaluar el modelo de regresión logística se realizó la prueba de *razón de verosimilitud*, que realiza una comparación entre el modelo ajustado y el modelo sin predictores. Para ello, se calculó la diferencia entre las desviaciones del modelo ajustado y el modelo nulo y los grados de libertad resultantes de la diferencia entre los grados de libertad de estos dos modelos, y con dicho valor se realizó una prueba chi-cuadrado cuyo resultado arrojó un estadístico de 127.74 con 39 grados de libertad (número de niveles existentes entre todas las variables) y un valor-p inferior a 0.001, el cual indica que el modelo ajustado en conjunto se ajusta significativamente mejor que el modelo nulo.

Por otro lado, se realizó una matriz de confusión, utilizando los mismos datos con los cuales se construyó el modelo, para clasificar las averías en mecánicas o eléctricas/electrónicas teniendo en cuenta los valores ajustados, por lo que si la probabilidad predicha era superior a 0.5 entonces la avería sería de tipo mecánico. Lo anterior se hizo como una forma de evaluar la eficiencia del modelo para clasificar correctamente las predicciones en los dos tipos de averías tenidos en cuenta, sin embargo, se aclara que la importancia del modelo ajustado era predecir la probabilidad de ocurrencia de una avería mecánica.

En la diagonal de la Tabla 3 se muestra el número de observaciones que se clasificaron de manera correcta según el tipo de avería, lo que se traduce en una eficiencia del 60.4% que puede interpretarse como aceptable.

Tabla 3. Matriz de confusión para las averías.

	Avería eléctrica	**Avería mecánica**
Avería eléctrica	445	430
Avería mecánica	312	688

Por último, la curva ROC (ver Figura 12), que mide el rendimiento global del modelo, señala que la probabilidad de clasificar correctamente es de 63.5%, cuyo valor es calculado a partir de la proporción estimada bajo la curva del gráfico.

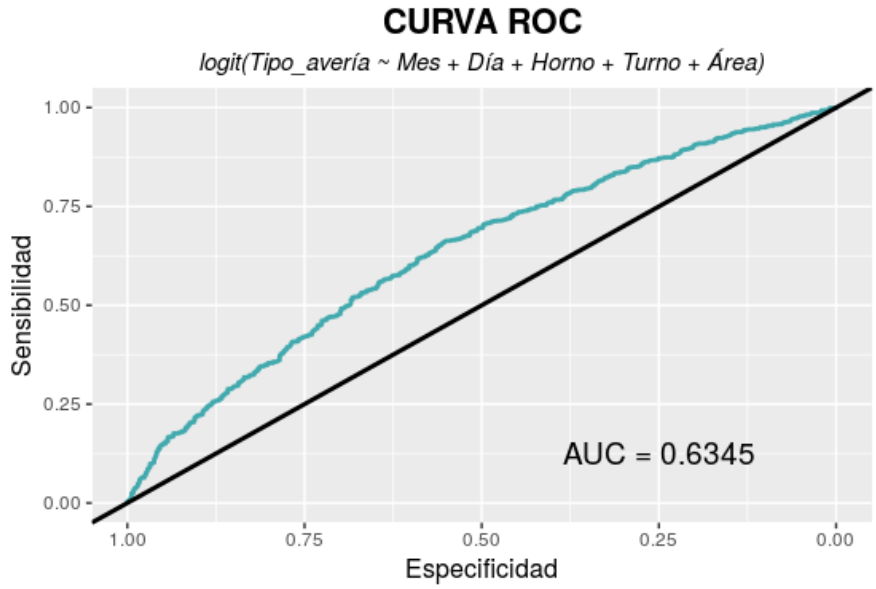

Figura 12. Curva ROC para el modelo de averías. Fuente: elaboración propia.

En la Figura 13 se muestra la distribución de los valores predichos y, por ejemplo, una de las predicciones es que, en el mes de noviembre, un día

martes, trabajando en el turno 1 y el horno 12 en el área de multiempaque hay una probabilidad de 67.19% de que ocurra una avería mecánica; mientras que para el mes de marzo si se está trabajando un lunes en turno 1 y en el horno 12 en el área individual hay una probabilidad de 34.85% de que ocurra una avería mecánica.

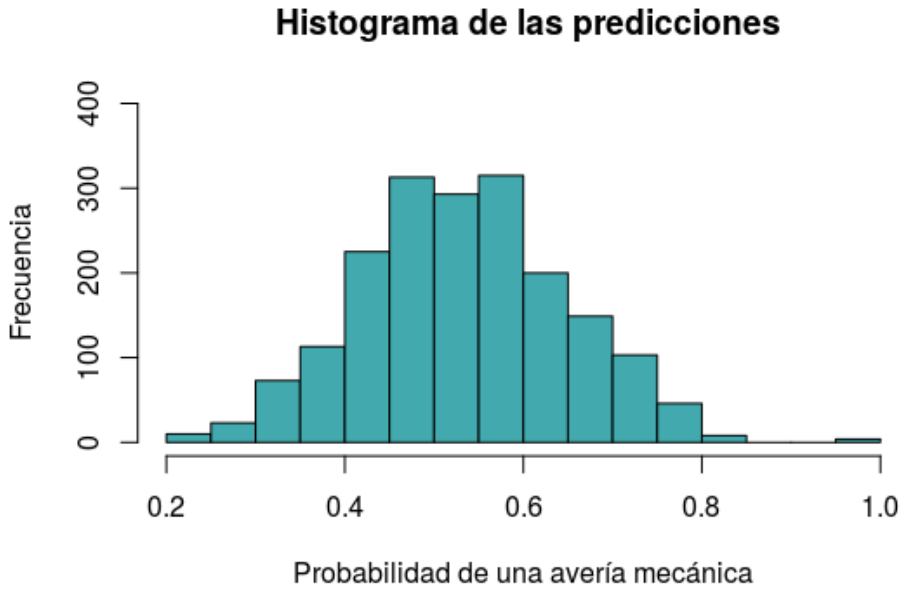

Figura 13. Histograma para los valores de probabilidad predichos. Fuente: elaboración propia.

5.2. Recorte/Reproceso de galleta

Como se mencionó, el reproceso o recorte es la cantidad de galleta no conforme que se genera en diferentes etapas del proceso por múltiples motivos, dicha galleta puede ser incorporada o no al proceso dependiendo de su estado. Para el análisis del reproceso de galleta se tuvieron en cuenta las variables descritas a continuación en la tabla Tabla 4.

Tabla 4. Descripción de variables para el análisis del reproceso.

Variable	Niveles	Unidad de medida
Horno	Z01, Z02, …, Z29	-
Mes	Enero, …, diciembre	-
Día	Lunes, …, domingo	-
Turno	1, 2 y 3	-
Tipo de reproceso	Ajuste de equipo, barredura, chicharrón, cobertura chocolate, con papel, cremada selección, dulce cremada, no conforme horno, orillo y simple empaque	-
Área de	Centro de empaque, cremadoras, embalaje,	-

reproceso	empacadoras verticales, empaque individual, multiempaque, recogida y recubridora chocolate	
Reproceso	-	Kilogramos

Si bien se tiene una meta para el porcentaje de reproceso en cada horno, al analizar los kilogramos de reproceso por horno durante todo el año 2018 se observa una alta variabilidad (ver Figura 14). Valores para el reproceso pueden ir desde 0.5 kg hasta 1000 kg, siendo los hornos 5 y 11 aquellos que presentan un reproceso medio mayor. Por otro lado, se analizaron las áreas del proceso donde se presentaba recorte con mayor frecuencia y se encontró que sólo en el área de multiempaque se concentra el 86.7% de reproceso (Ver Figura 15). Por tal razón, lo adecuado era concentrarse en dicha área, así que el análisis predictivo estuvo enfocado allí en particular.

Debido entonces a la alta variabilidad observada y al foco puesto en el área de multiempaque, se depuraron los datos restringiendo aquellas observaciones con reproceso en el área de multiempaque cuyas cantidades de recorte fueran inferior a los 56.3 kilogramos; valor por el cual las observaciones por encima de este número son consideradas atípicas, esto a partir de un análisis gráfico del boxplot de la variable respuesta en donde el límite superior del bigote correspondía a dicho valor. También se priorizaron los tipos de reproceso: el recorte con papel, por barredura, simple empaque, no conforme horno y por ajuste de equipo, que representan el 93.8% del total de reproceso, y porque el reproceso por dulce cremada, chicharrón, orillo, cremada selección y cobertura chocolate se presenta solamente en ciertas referencias de galletas. De esta forma, se excluyeron 12070 observaciones de 155844 que habían originalmente, esto es, un 9.5% del total de observaciones.

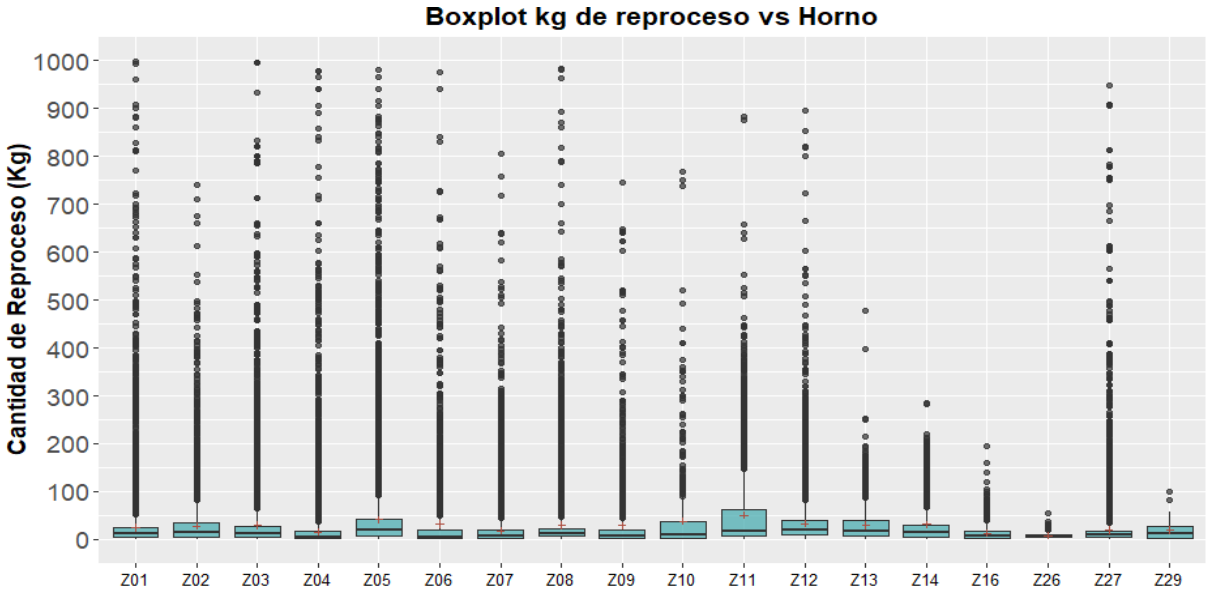

Figura 14. Boxplot para los kilogramos de reproceso en función del horno. Fuente: elaboración propia.

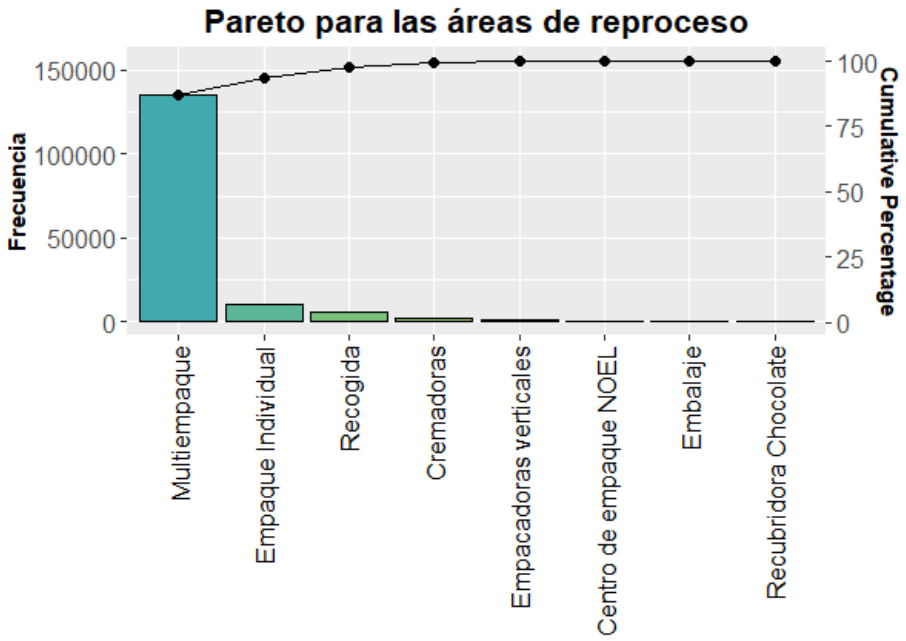

Figura 15. Pareto de la frecuencia de reproceso en las diferentes áreas. Fuente: elaboración propia.

Posterior a la depuración de los datos, se realizó nuevamente un análisis exploratorio. Primero se analizó el reproceso por mes (Ver Figura 16) y se observó un leve incremento de los kilogramos de recorte durante los meses de junio y julio, en los cuales en promedio se registran 14.7 y 14.5 kilogramos de reproceso, respectivamente. Luego de estos dos meses, los kilogramos de recorte tienden a estabilizarse nuevamente, así que en general no pareciera haber diferencia alguna durante el año, en cuanto a reproceso se refiere. Resultados similares se observan en la Figura 17 en donde se comparan los kilogramos de reproceso de acuerdo al día de la semana. No parece que el día de la semana influya en el reproceso.

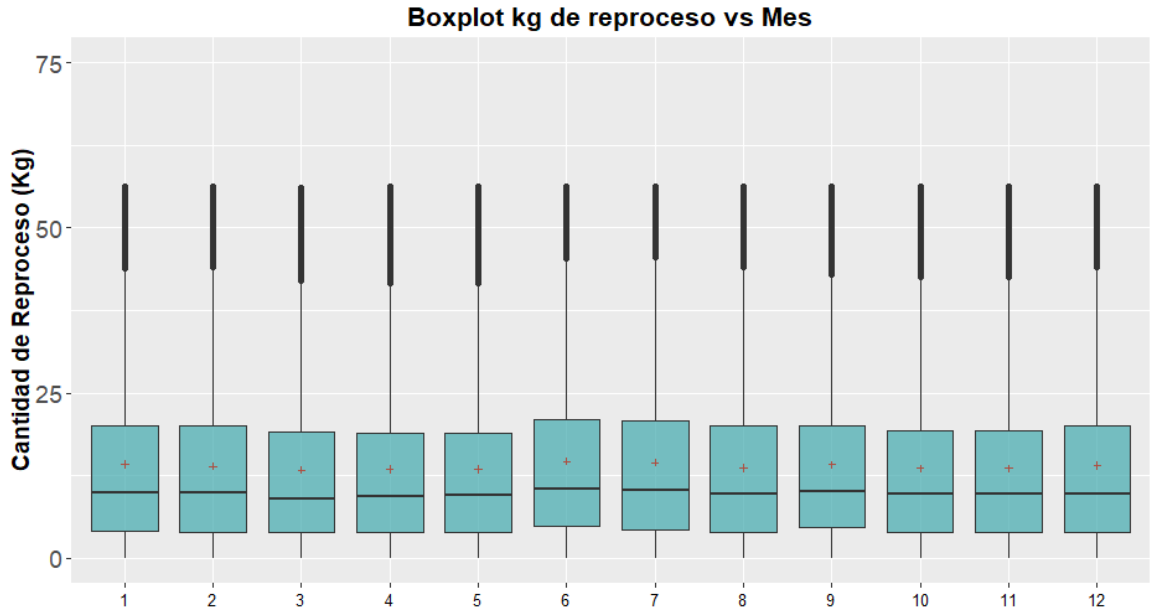

Figura 16. Boxplot de los kilogramos de reproceso en función del mes. Fuente: elaboración propia.

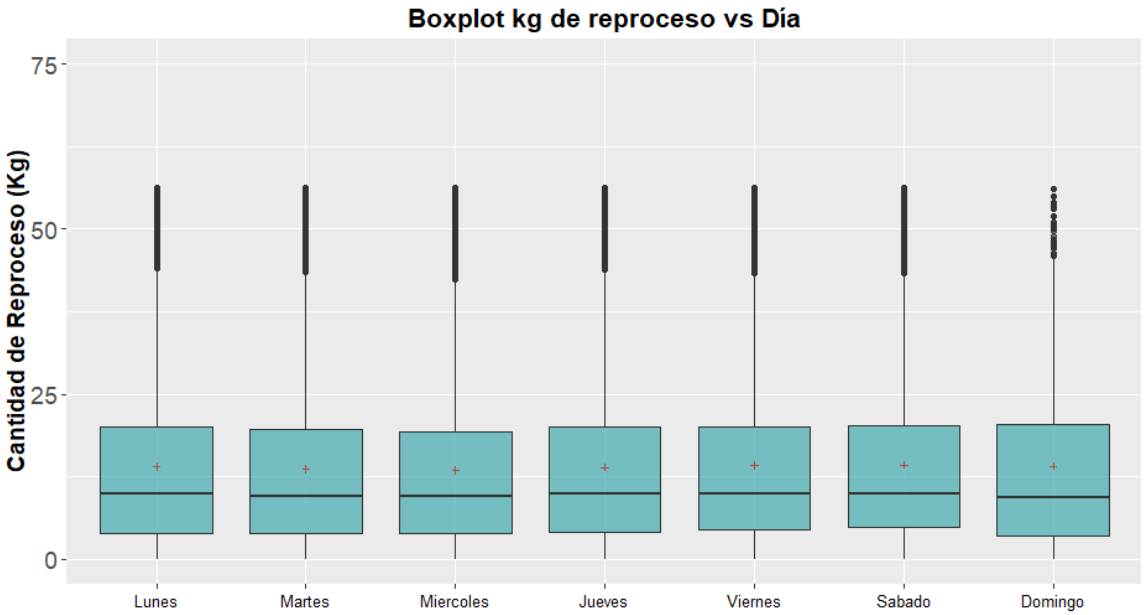

Figura 17. Boxplot de los kilogramos de reproceso según el día de la semana. Fuente: elaboración propia.

En cuanto al turno, se tiene que el turno 1 registra en promedio 13.9 kilogramos de recorte, el turno 2 presenta un reproceso levemente mayor (14.2 kg), mientras que el turno 3 tiene un recorte promedio de 13.7 kilogramos; la diferencia entre los tres turnos no parece ser significativa.

Por otro lado, se analizó el recorte dependiendo del tipo de reproceso (Ver Figura 18) y sí se evidenciaron diferencias que pudiesen ser significativas a la hora de explicar la variable respuesta. Se reporta mayor sobrepeso si éste es debido a ajustes en equipos, lo que parece lógico dado que se alteran ciertas condiciones de operación y esto ocasiona mayor cantidad de galleta no conforme. El reproceso por no conformidad de horno registra un recorte de 17.2 kg en promedio, mientras que el reproceso por barredura (galleta que cae al piso) es el tipo de recorte con menor cantidad reportada (9.3 kilogramos).

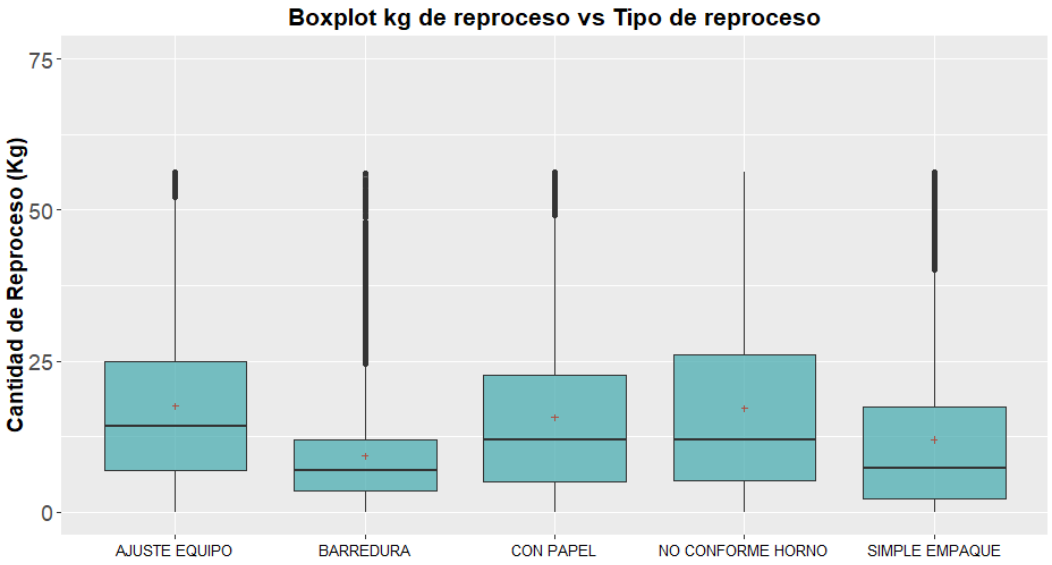

Figura 18. Boxplot de los kilogramos de reproceso en función del tipo de reproceso. Fuente: elaboración propia.

En la compañía son conscientes de que el reproceso en el horno 5 es muy elevado y por eso han empezado a implementar planes de mejoramiento para disminuirlo. En la Figura 19 se muestra de forma clara que uno de los hornos con mayor recorte durante el 2018 fue el 5, con un reproceso promedio de 18.9 kilogramos, superado solamente por el horno 12 (el horno con mayor producción de la compañía), cuyo reproceso medio es de 21.2 kg. En la gráfica también se observa que el horno pudiese ser una de las variables que mejor explique el reproceso dada la marcada diferencia existente entre los hornos en cuanto a kilogramos de recorte.

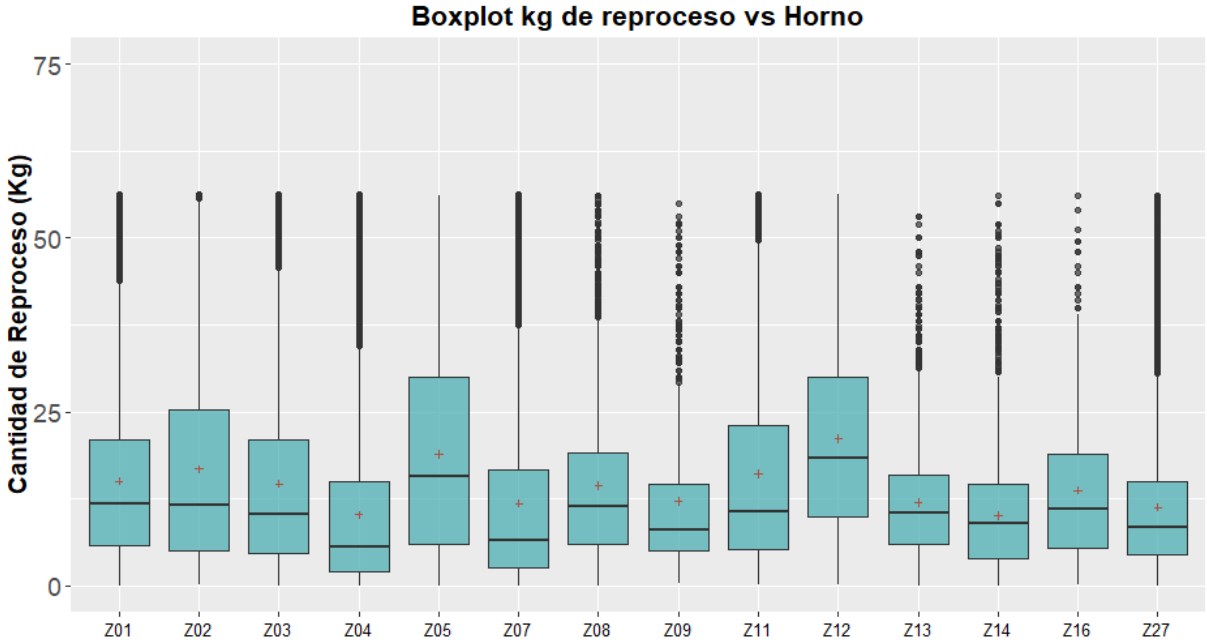

Figura 19. Boxplot de los kilogramos de reproceso en función del horno. Fuente: elaboración propia.

Para complementar el análisis exploratorio se construyó el árbol de regresión mostrado en la Figura 20, el cual da una idea de las predicciones que se realizarán utilizando otros modelos predictivos de regresión y las variables con mayor importancia. Cada uno de los rectángulos representa un nodo del árbol con una predicción de los kilogramos de reproceso promedio y también se muestra la proporción de casos a las que se les atribuyó dicha predicción.

Según la Figura 20 las variables más importantes a la hora de explicar el sobrepeso serían el horno y el tipo de reproceso, en donde el 47% de las observaciones tendrían un reproceso promedio de 11 kilogramos siempre y cuando se estuviera mirando el recorte para los hornos Z01, Z07, Z09, Z13, Z14 Y Z27, mientras que, si se trabaja en los hornos Z02, Z05 y Z12 y además se está analizando el recorte por ajuste de equipo, con papel, no conforme horno o simple empaque, se tendría un reproceso medio de 21 kilogramos , lo que es coherente con lo analizado previamente en el boxplot de la cantidad de reproceso por horno en el cual se observaron mayores valores de recorte para los hornos Z05 y Z12.

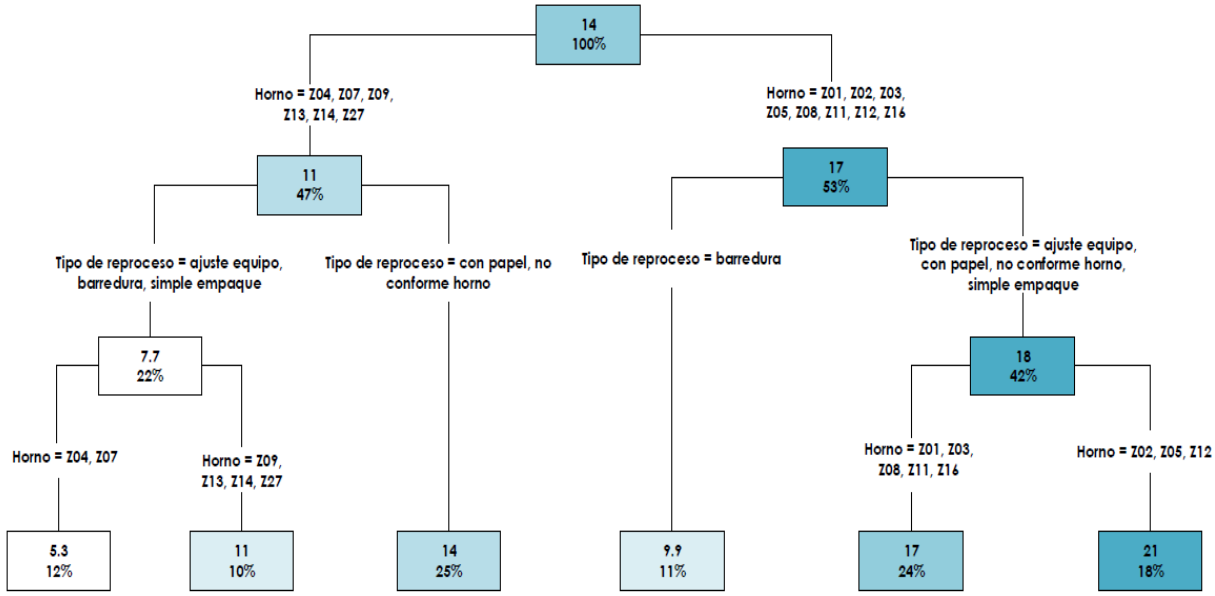

Figura 20. Árbol de regresión para el reproceso en el área de multiempaque. Fuente: elaboración propia.

5.2.1. Bosques aleatorios y Máquinas de soporte vectorial

En el análisis exploratorio se utilizó un árbol de regresión para complementar el análisis descriptivo y no se usó como un método predictivo, dado que usar sólo un árbol puede crear un sobreajuste y predicciones muy variables cada vez que se ejecute el algoritmo, ya que cada ejecución crea un árbol diferente que puede ser parecido o no al primer árbol creado. Se observó entonces la distribución de la variable respuesta para evaluar la posibilidad de implementar un modelo de regresión lineal múltiple, sin embargo, como

se observa en la Figura 21 el reproceso no sigue una distribución normal y no se optó por transformar la variable respuesta.

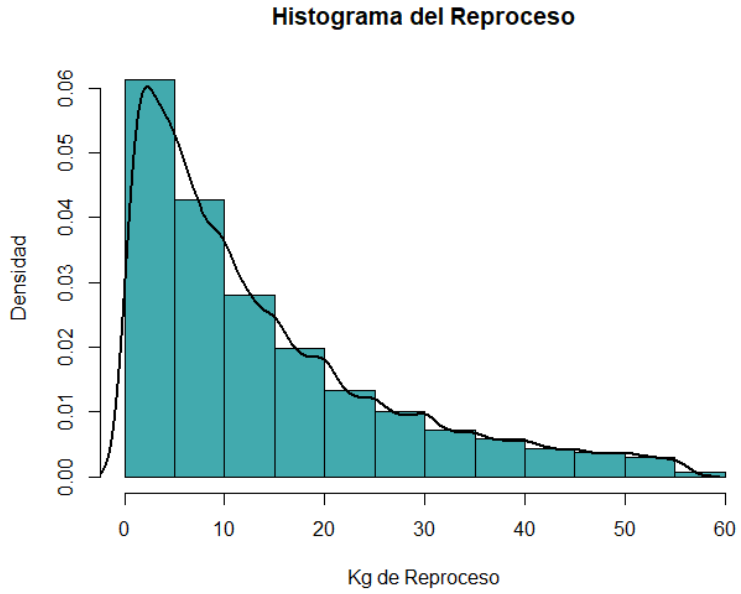

Figura 21. Distribución del reproceso en el área de multiempaque. Fuente: elaboración propia.

Para la predicción del reproceso en el área de multiempaque se decidió utilizar la técnica de bosques aleatorios en la cual puede controlarse el número de árboles construidos y al final elegir el mejor árbol para la predicción. También se utilizaron las máquinas de soporte vectorial cuyo algoritmo ha demostrado alta eficiencia a la hora de resolver problemas de regresión y clasificación, y al final los resultados fueron comparados con los del bosque aleatorio para elegir el mejor modelo y realizar las predicciones.

En ambos casos, bosques aleatorios y máquinas de soporte vectorial, los datos iniciales fueron divididos en dos conjuntos: el 80% de los datos conformaron el conjunto de datos de entrenamiento, es decir, los datos con los que el algoritmo aprenderá por sí mismo, mientras que el 20% de los datos restantes conformaron el conjunto de datos de prueba con los cuales se testearon y validaron los modelos una vez fueron creados. Esta partición es respaldada por expertos en machine learning, como Harrington (2012) quien en su libro titulado "*Machine Learning in action*" habla de cómo el conjunto de entrenamiento debe ser representativo acogiendo cerca del 70% - 80% de los datos.

Es necesario aclarar que inicialmente fueron construidos los modelos del bosque aleatorio y las máquinas de soporte vectorial teniendo en cuenta los datos iniciales del área de multiempaque incluidos los puntos atípicos, pero debido al bajo desempeño de los modelos de acuerdo a las validaciones realizadas, se decidió ejecutar nuevamente los algoritmos teniendo en cuenta los datos del área de multiempaque para los cuales el reproceso es inferior a 56.3 kilogramos; valor por el cual las observaciones por encima de este número son consideradas atípicas. De allí, el análisis exploratorio volvió a

repetirse con los resultados que anteriormente se mencionaron en dicho apartado.

Para la construcción del bosque aleatorio se utilizaron 500 árboles de regresión y se encontró que las variables más importantes a la hora de explicar el reproceso en el área de multiempaque fueron el horno y el tipo de reproceso, lo que concuerda con el análisis exploratorio realizado luego de la depuración de los datos iniciales y al árbol de regresión modelado. Por su parte, el turno no pareció ser tan importante a la hora de explicar el recorte, sin embargo, se tomó en consideración para la realización de las predicciones.

Por otro lado, se planteó un modelo de máquinas de soporte vectorial considerando las mismas variables del bosque aleatorio y utilizando el mismo conjunto de datos. Por defecto se utilizó un costo igual a 1, pero dicho parámetro podría optimizarse para mejorar los resultados del modelo. La función Kernel utilizada fue la radial con parámetros gamma igual a 0.02 y épsilon igual a 0.1, según lo estimó conveniente el algoritmo a partir del conjunto de datos de entrenamiento. En total fueron necesarios 91569 vectores de soporte, que son demasiados, pero tiene sentido dada la alta variabilidad de los datos.

5.2.2. Evaluación del modelo y predicciones

Para la validación de los modelos se utilizó el conjunto de datos de prueba y como medida de evaluación y comparación se utilizó el error cuadrático medio y la correlación. Estos resultados se muestran en la Tabla 5 y en ambos casos se obtuvieron bajas correlaciones, pero hay que tener en cuenta que se hizo un modelo para la predicción de una variable cuantitativa en donde todas las variables independientes eran cualitativas y con múltiples niveles, lo que disminuye eficiencia y calidad.

Tabla 5. Error cuadrático medio y correlación para los modelos de reproceso ajustados.

Modelo	Error cuadrático medio	Correlación
Bosques aleatorios	132.55	0.48
Máquinas de soporte vectorial	149.43	0.40

Para las predicciones finales se eligió el modelo de bosques aleatorios por su mayor correlación y menor error cuadrático medio. De esta manera, un ejemplo de predicción podría ser que, si la compañía se encuentra produciendo en el mes de junio, un viernes durante el turno 3 y en el horno 12, va a tener 15.71 kilogramos de reproceso por barredura en el área de multiempaque.

5.3. Sobrepeso Mix de la Galleta Saltín Fit taco x 5 en el Horno 12.

El sobrepeso mix es un indicador porcentual que permite saber si se produjo más kilogramos de lo esperado. Si este valor se encuentra por encima de la meta propuesta por la compañía, entonces se incurre en pérdidas debido a que se está gastando más materia prima de la presupuestada. En la Tabla 6 se muestra con mayor detalle la definición y unidad de medida de las variables tenidas en cuenta para la modelación del porcentaje de sobrepeso mix.

Tabla 6. Descripción de variables para el análisis del sobrepeso mix.

Variable	Niveles	Unidad de medida
Mes	Febrero, …, diciembre	-
Día	Lunes, …, domingo	-
Turno	1, 2 y 3	-
Peso de 10 galletas	-	Gramos
Resistencia promedio de la galleta	-	Gramos
Ancho de la galleta	-	mm
Calibre	-	unidades
PH	-	(adimensional)
Humedad	-	%
Sobrepeso Mix	-	Porcentaje

En la Figura 22 se muestra el porcentaje de sobrepeso mix de la galleta Saltín Fit taco x 5 en el horno 12 desde febrero hasta diciembre de 2018 y se observa una alta variabilidad, siendo los meses de febrero y diciembre los de mayor porcentaje de sobrepeso promedio con 3.36% y 3.35%, respectivamente. Sin embargo, el mes de mayo es el que presenta mayor variabilidad alcanzando un valor máximo de 7.5% y un valor mínimo de sobrepeso de -3.38%.

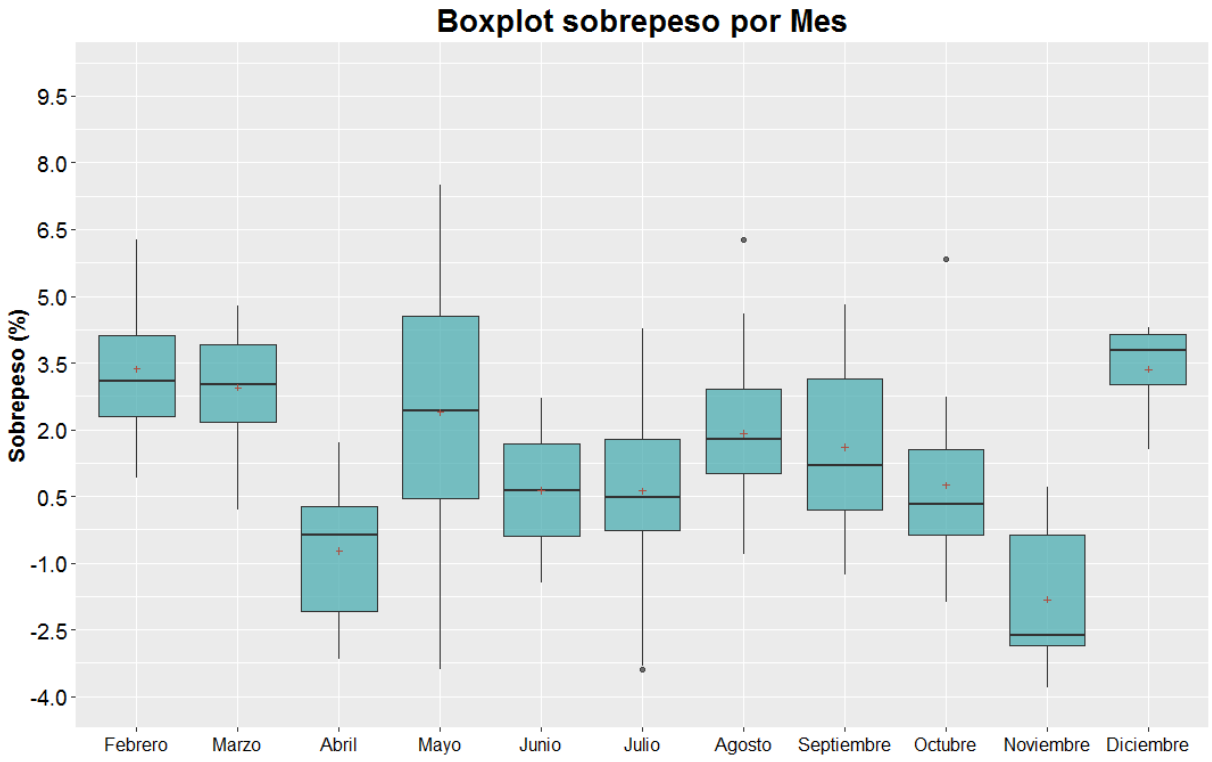

Figura 22. Boxplot para el porcentaje de sobrepeso mix en función del mes. Fuente: elaboración propia.

También, como se mencionó anteriormente, es probable que los indicadores varíen de un turno a otro, ya que en la producción los colaboradores suelen cambiar algunos métodos o ajustar condiciones de los equipos según sea su preferencia; por tanto, en este trabajo también pudo analizarse si el turno realmente influía en el sobrepeso de la referencia estudiada. En la Figura 23 se observa que el turno parece influir significativamente en el sobrepeso, por lo que esta variable puede ser un buen predictor en los modelos planteados más adelante, siendo el turno 3 (turno que va desde las 9:40 p.m. hasta las 5:40 a.m.) aquel que trabaja con un mayor porcentaje de sobrepeso; una afirmación que también respalda la Dirección de Producción a partir del seguimiento realizado por medio de los indicadores calculados.

Por otro lado, no parecen haber diferencias significativas en el porcentaje de sobrepeso de la galleta Saltín Fit taco x 5 en cuanto al día de la semana se refiere (ver Figura 24). Aunque pareciera que la variabilidad aumenta al transcurrir la semana, el valor medio del sobrepeso no parece mostrar grandes cambios, exceptuando los días sábados y domingos, en los cuales se tiene un valor de medio para el sobrepeso de 3.32% y 3.06%, respectivamente.

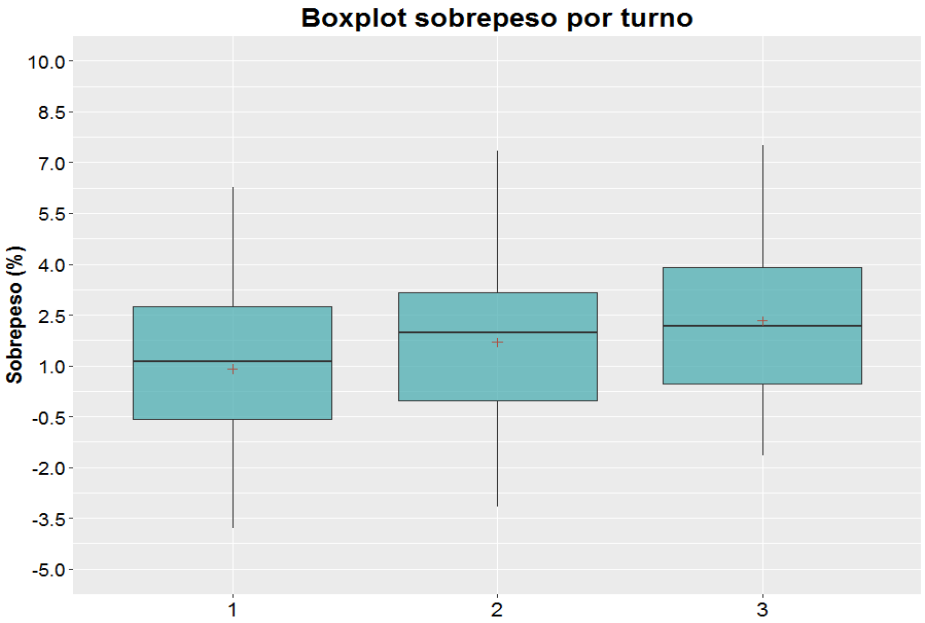

Figura 23. Boxplot para el porcentaje de sobrepeso mix en función del turno. Fuente: elaboración propia.

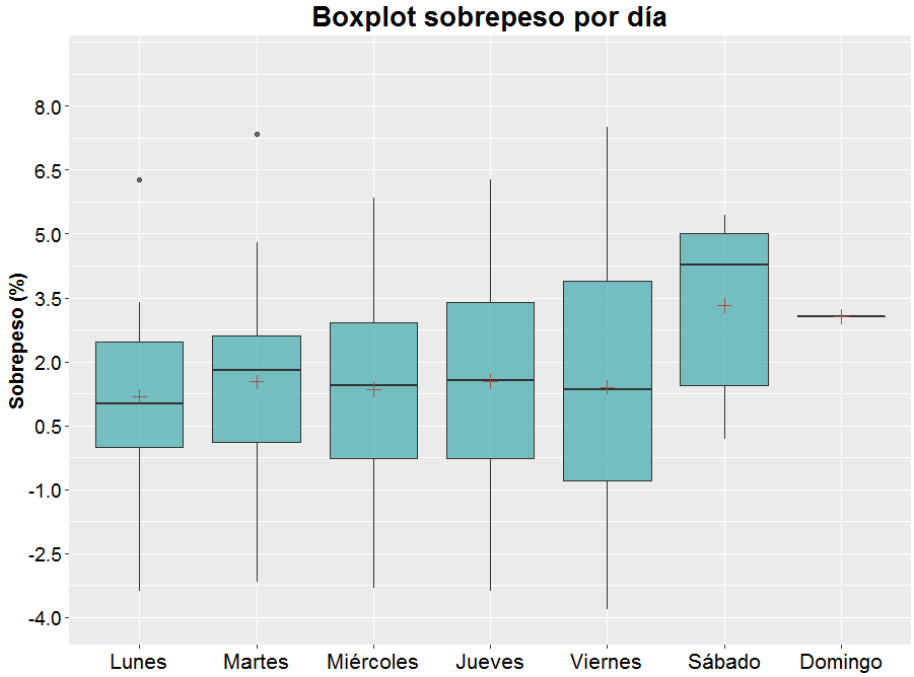

Figura 24. Boxplot para el porcentaje de sobrepeso mix según el día de la semana. Fuente: elaboración propia.

Una de las variables independientes tenidas en cuenta para el análisis predictivo del sobrepeso fue la resistencia promedio de la galleta. Este valor es obtenido a partir del peso que la galleta puede soportar sin quebrarse. En la Figura 25 se muestra el comportamiento de la resistencia promedio de la galleta en función del mes, y se observa que durante el mes de julio se registran mayores valores de resistencia, pero no es claro si esto pudo influir en el comportamiento del sobrepeso para ese mes en particular. Tampoco

puede inferirse si la alta variabilidad del sobrepeso observada durante el mes de mayo tenga que ver con los valores de resistencia promedio registrados en dicho mes, así lo que las respuestas a estos cuestionamientos tendrán que ser abordadas más adelante durante el análisis del modelo predictivo planteado. De momento, a partir del gráfico podría decirse que la resistencia promedio puede ser una variable significativa a la hora de explicar el sobrepeso, ya que se observan diferencias y esto puede que realmente esté afectado el peso de la galleta.

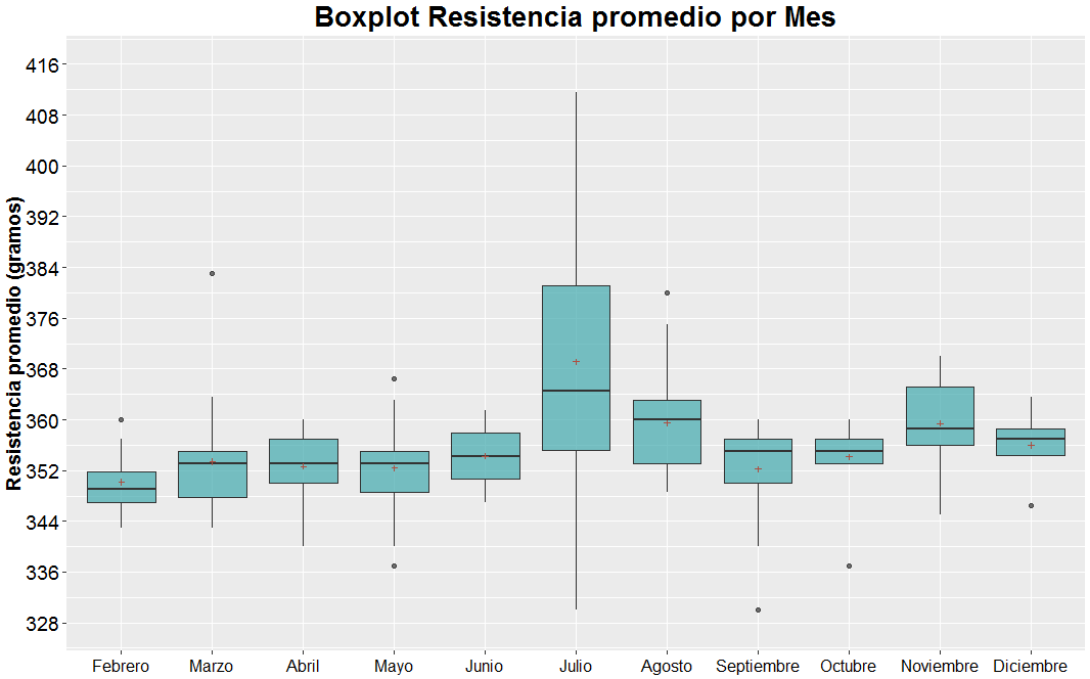

Figura 25. Boxplot para la resistencia promedio de la galleta en función del mes. Fuente: elaboración propia.

El diagrama de dispersión en 3D de la Figura 26 muestra la relación entre el PH, la humedad de la galleta y el sobrepeso mix por turno. No se observan relaciones directas lineales entre estas tres variables más el turno, pero se incluirán términos cuadráticos en algunos modelos planteados y se evaluará su significancia. De manera descriptiva no parece que hubiese relación entre el PH y el sobrepeso y de igual forma no se observan relaciones entre la variable respuesta y la humedad.

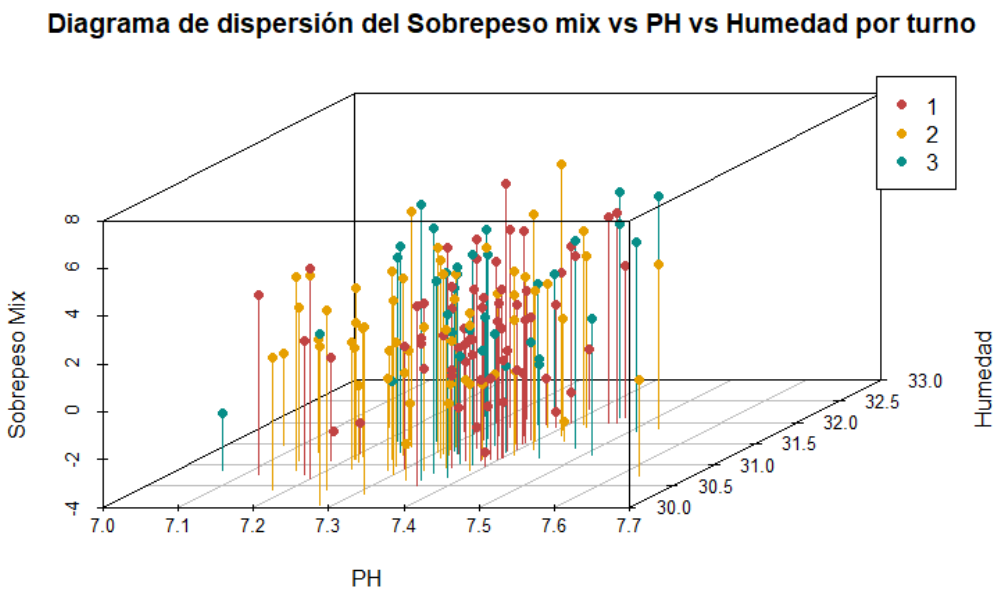

Figura 26. Diagrama de dispersión del sobrepeso mix vs PH vs humedad por turno. Fuente: elaboración propia.

5.3.1. Regresión lineal, modelos GAMLSS, bosques aleatorios y máquinas de soporte vectorial

Como el objetivo es predecir el porcentaje de sobrepeso mix, lo usual es pensar en un modelo adaptado a datos que toman valores desde cero hasta uno. Sin embargo, el sobrepeso 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 (como el modelo de regresión beta) y mejor se optó por utilizar otras técnicas dependiendo de la distribución de los datos.

De manera descriptiva se realizaron pruebas gráficas y analíticas para verificar la distribución del sobrepeso mix y, como se observa en la Figura 27, la variable respuesta sigue una distribución normal. Esto se hizo para evaluar la posibilidad de implementar un modelo de regresión lineal múltiple, ya que una de las condiciones que debe cumplirse es la normalidad de la variable respuesta. Por otro lado, se realizó una matriz de dispersión con el fin de observar posibles relaciones lineales entre cada par de variables cuantitativas del conjunto de datos (ver Figura 28) y se observó que no existen altas correlaciones entre las variables independientes, ni entre éstas y la variable respuesta, aunque se evidencian correlaciones no tan pequeñas entre el peso de 10 galletas y el calibre con el sobrepeso mix.

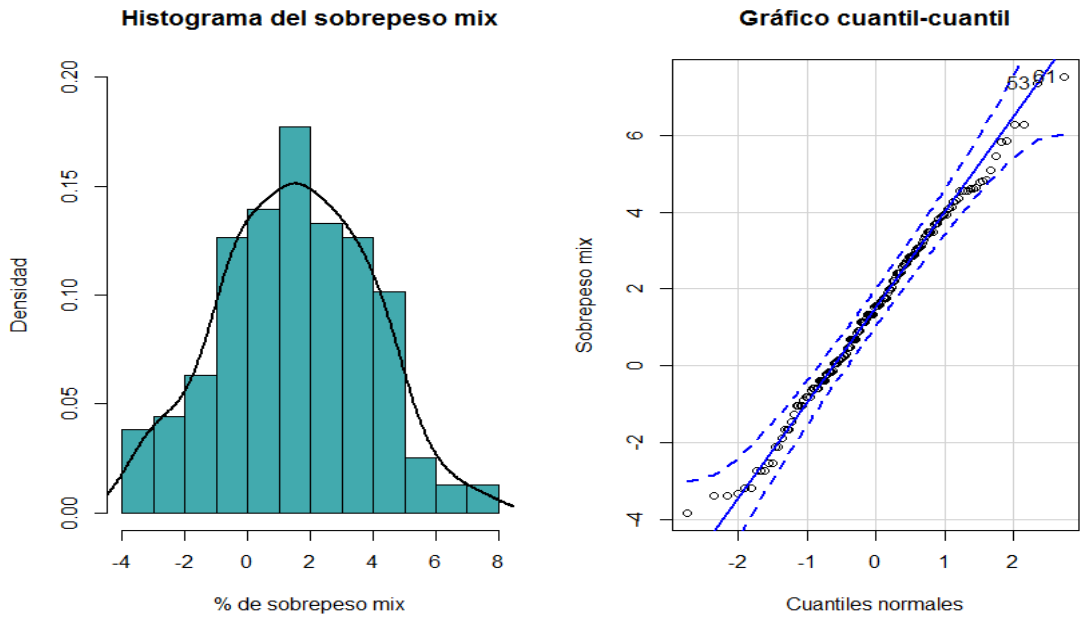

Figura 27. Distribución del porcentaje sobrepeso mix. Fuente: elaboración propia.

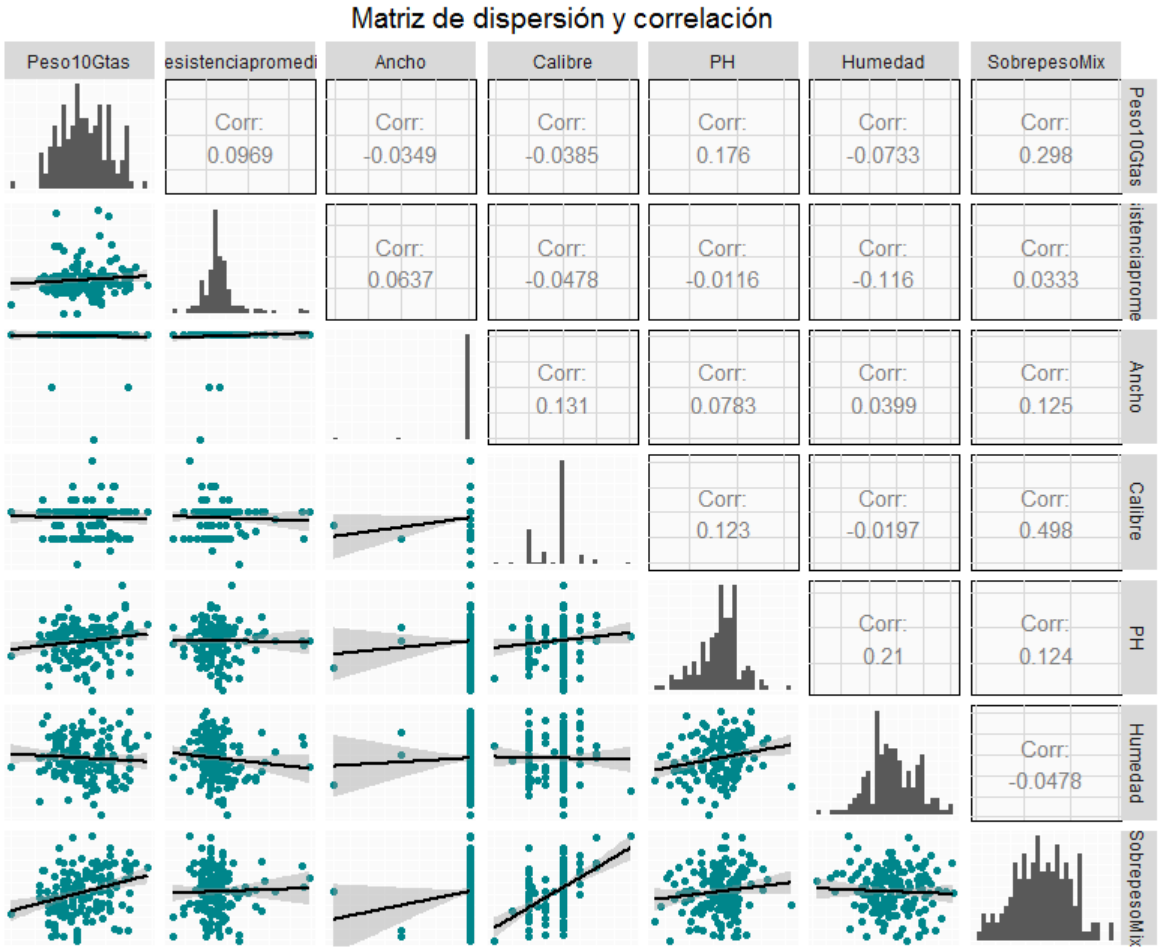

Figura 28. Matriz de dispersión y correlación para las variables del análisis del sobrepeso mix. Fuente: elaboración propia.

- **Modelo de regresión lineal múltiple**

Se planteó un modelo de regresión lineal múltiple considerando todas las variables para el sobrepeso descritas previamente y se hizo una selección de variables utilizando el *criterio de información de Akaike* (1974) a través de una combinación de los métodos de selección *forward* y *backward*, donde no resultaron significativas las variables día de la semana, ancho de la galleta, PH y humedad, mientras que las más significativas fueron el mes y el calibre de la galleta. Finalmente, se obtuvo un modelo con la siguiente estructura:

$$\begin{aligned} \%\, Sobrepeso\ Mix = {} & \beta_0 + \beta_1 Mes_i + \beta_2 Turno_j + \beta_3 Peso10galletas + \\ & \beta_4 Resistenciapromedio + \beta_5 Resistenciapromedio^2 + \\ & \beta_6 Calibre + \beta_7 Calibre^2 \end{aligned} \qquad (9)$$

siendo $i = 1, \ldots, 158$ y recordando que para las variables categóricas se genera un coeficiente para cada uno de los niveles a excepción del nivel de referencia. Por ejemplo, para la variable *turno* se tendrán dos coeficientes estimados y el *turno 1* será el nivel de referencia. Los coeficientes para cada una de las variables del modelo ajustado se muestran a continuación en la Tabla 7. En el caso del predictor *peso de 10 galletas*, si el resto de variables permanecen constantes, por cada unidad de peso que aumenten las 10 galletas pesadas, el porcentaje de sobrepeso mix se incrementa en promedio 1.17 unidades. Para el caso del turno, se tiene que el turno 3 obtiene en promedio 1.02 unidades de sobrepeso más que el turno 1 (nivel de referencia) y esto concuerda con el análisis exploratorio en donde se observó que el turno sí podría ser significativo a la hora de explicar el sobrepeso, especialmente el turno 3 que mostró los mayores valores de esta variable.

Tabla 7. Coeficientes del modelo de regresión lineal múltiple estimado para el sobrepeso mix.

Variable	Coeficiente estimado	Variable	Coeficiente estimado
Mes - marzo	-0.679	Mes - diciembre	-5.820
Mes – abril	-3.376	Turno 2	0.562
Mes – mayo	-1.536	Turno 3	1.022
Mes – junio	-2.451	Peso10Gtas	1.174
Mes – julio	-3.214	Resistencia promedio	0.514
Mes – agosto	-1.373	Resistencia promedio^2	0.001
Mes – septiembre	-0.890	Calibre	-46.08
Mes – octubre	-2.107	Calibre^2	0.662
Mes – noviembre	-4.735		

- **Modelos GAMLSS**

También, se decidió aplicar modelos GAMLSS para explicar el porcentaje de sobrepeso en función de las mismas variables independientes tenidas en cuenta hasta ahora.

Se ajustaron las cuatro distribuciones estadísticas que mejor explicaron el comportamiento del sobrepeso sin incluir las variables independientes, teniendo en cuenta que dichas distribuciones tienen como dominio los números reales. En la Figura 29 se muestra el histograma y la densidad de cada una de las distribuciones ajustadas: normal, normal generalizada o power exponential, power exponential tipo 2 y skew-normal tipo 2; en ese orden. Se observa que la distribución normal describe de manera adecuada el comportamiento del sobrepeso mix usando dos parámetros (μ y σ), mientras que las otras tres distribuciones lo hacen por medio de tres parámetros (μ, σ y ν).

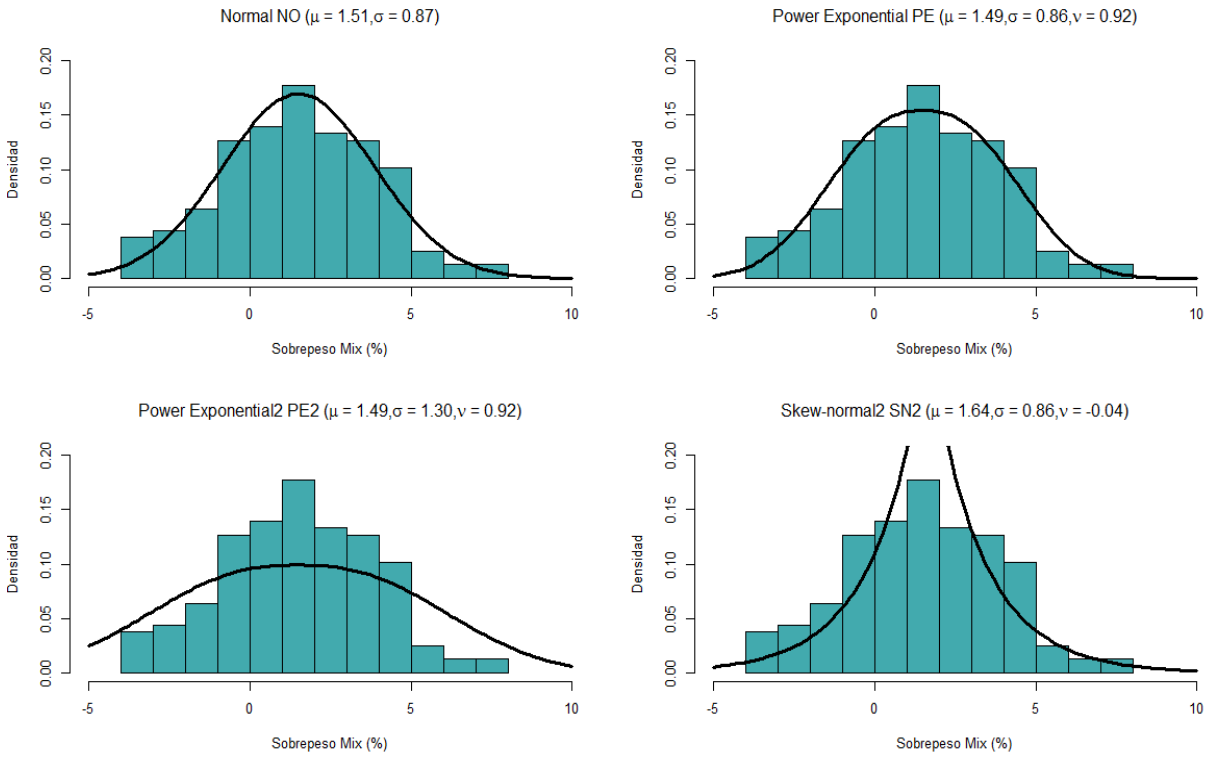

Figura 29. Histograma para el sobrepeso mix con las cuatro densidades de probabilidad que mejor se ajustan a la variable respuesta. Fuente: elaboración propia.

Se realizaron cuatro modelos, uno por cada distribución ajustada y considerando términos lineales y cuadráticos de las variables cuantitativas. Además, se realizó una selección de variables con el objetivo de elegir las variables significativas para el modelo.

Para realizar la comparación entre los modelos ajustados bajo la metodología GAMLSS se utilizó el *Akaike information criterion* (AIC) tomando un valor de penalidad k igual al $log(n)$, donde n es igual al número de observaciones. También se analizó el worm plot (Buuren & Fredriks, 2001) para la selección del mejor modelo. En la Tabla 8 se presenta el AIC de los cuatro modelos ajustados y se observa que los modelos con variable respuesta power exponential y power exponential 2 tienen menores valores AIC, sin embargo, al analizar el worm plot para cada uno en la Figura 30, se observa que el modelo con variable respuesta normal tiene un mejor desempeño, pues sus valores residuales no invaden las hipérbolas, lo que indica un buen ajuste. Además, debe tenerse en cuenta que este modelo utiliza un menor número de parámetros comparado con los otros tres. Así que, bajo el principio de parsimonia, el criterio AIC y lo observado en el worm plot, se decide seleccionar el modelo con variable respuesta normal.

Tabla 8. AIC para los modelos de las cuatro mejores distribuciones ajustadas.

Modelo (GAMLSS)	Distribución	Grados de libertad	AIC
1	NO	30	702.21
2	PE	43	657.50
3	PE2	41	670.28
4	SN2	25	4035.06

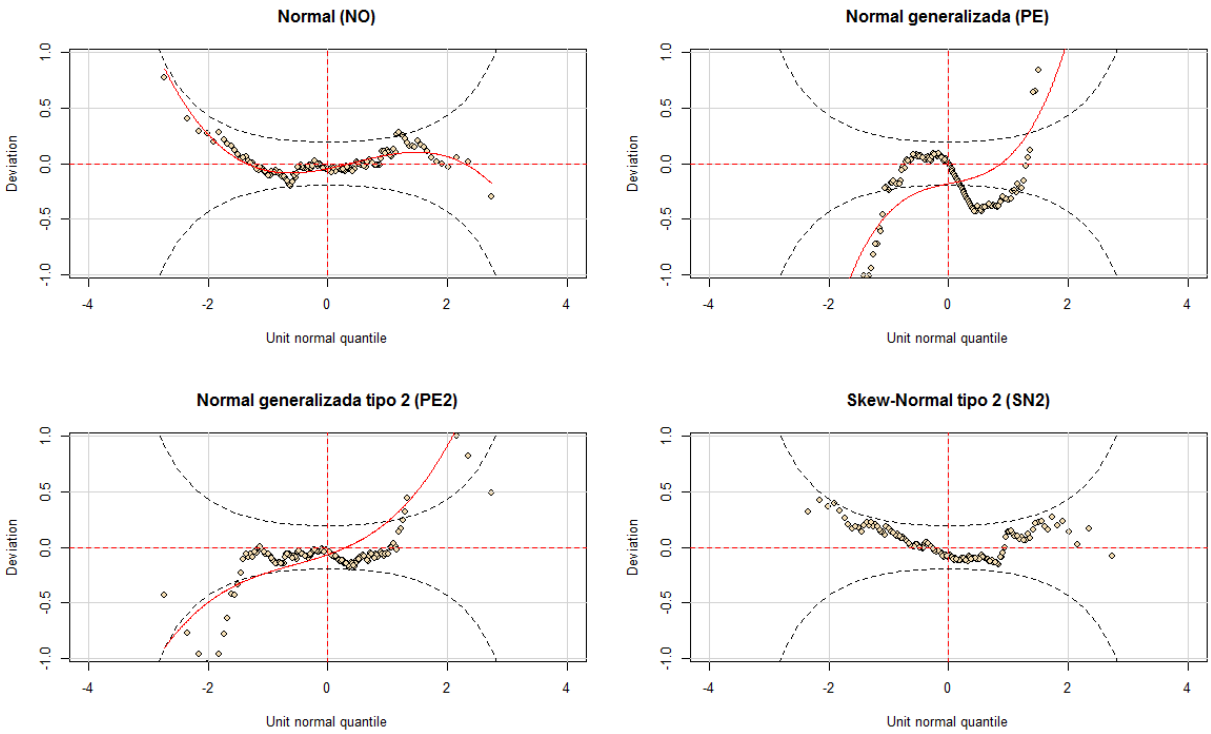

Figura 30. Worm plot para cada uno de los cuatro modelos ajustados. Fuente: elaboración propia.

En la Tabla 9 se presentan los parámetros estimados para el modelo con variable respuesta normal (NO). Cada una de las variables es significativa para los para los parámetros μ y σ teniendo en cuenta un α igual a 0.05 y recordando que se hizo una previa selección en donde las variables día, ancho, PH y humedad fueron eliminadas.

Tabla 9. Parámetros estimados para el modelo ajustado con distribución normal NO.

Modelo para μ			
Variable	**Coeficiente estimado**	**Variable**	**Coeficiente estimado**
Intercepto	998.3	Mes – noviembre	-5.378
Mes - marzo	-0.583	Mes - diciembre	-0.652
Mes – abril	-3.471	Turno 2	0.850
Mes – mayo	-1.506	Turno 3	1.137
Mes – junio	-2.888	Peso10Gtas	1.187
Mes – julio	-3.516	Resistencia promedio	-0.476
Mes – agosto	-1.620	Resistencia promedio^2	0.001
Mes – septiembre	-8.962	Calibre	-56.43
Mes – octubre	-2.266	Calibre^2	0.805
Modelo para $Log(\sigma)$			
Variable	**Coeficiente estimado**	**Variable**	**Coeficiente estimado**
Intercepto	6.375	Mes – agosto	0.007
Mes - marzo	-0.269	Mes – septiembre	-0.114
Mes – abril	-0.361	Mes – octubre	0.162
Mes – mayo	0.058	Mes – noviembre	0.302
Mes – junio	0.169	Mes - diciembre	-3.128
Mes – julio	-0.187	Resistencia promedio	-0.017

- **Bosques aleatorios y máquinas de soporte vectorial**

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 y se encontró que la variable más importante en la predicción del sobrepeso mix era el mes, seguido por el 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 previamente ajustado.

En el caso del modelo de máquinas de soporte vectorial se buscó optimizar los hiperparámetros para lograr un mejor desempeño del modelo dando como resultado un parámetro para el costo igual a 5, sin embargo, en la evaluación del modelo se observaron mejores resultados con un parámetro de costo igual a 1. En cuanto a la importancia de las variables, el modelo

máquinas de soporte vectorial arrojó los mismos resultados que los dos modelos ajustados anteriormente.

5.3.2. Evaluación del modelo y predicciones

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 10). El R^2 de dicho modelo fue de 0.56 y el R^2 ajustado fue de 0.51.

Tabla 10. Error cuadrático medio y correlación para los modelos ajustados del sobrepeso mix.

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

Los supuestos del modelo de regresión lineal también fueron validados: los predictores son independientes y no existen problemas de multicolinealidad, lo que apoya lo observado en la matriz de dispersión entre las variables independientes. Los residuos se distribuyen de forma normal, para ello se emplearon pruebas gráficas como el gráfico cuantil-cuantil y el histograma, y pruebas analíticas como la *shapiro-wilk*. También se verificó que se cumpliera con el supuesto de varianza constante de los residuos, en donde al representar los residuos frente a los valores ajustados por el modelo no se observó ningún patrón específico, como se observa en la Figura 31.

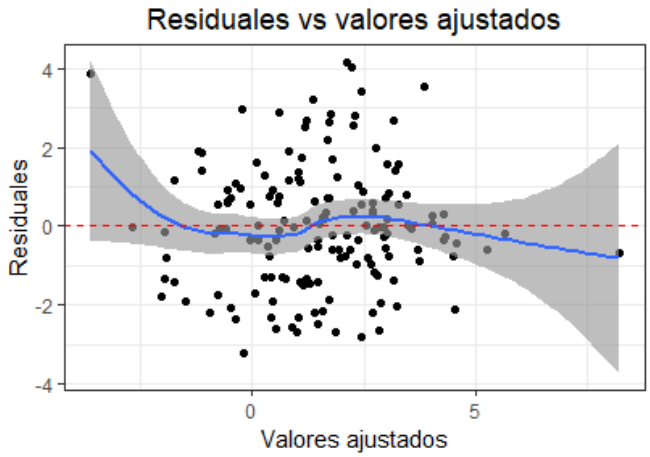

Figura 31. Residuales vs valores ajustados. Fuente: Elaboración propia.

Habiendo evaluado y comparado los modelos, se realizaron las predicciones para el porcentaje de sobrepeso mix de la galleta Saltín Fit del horno 12

utilizando el modelo de regresión lineal múltiple ajustado. Un ejemplo de predicción sería: si se produce esta galleta en el mes de febrero durante el turno 3, y además el peso de 10 galletas luego de su salida del horno fue igual a 60.6 gramos, la resistencia de una galleta fue de 360 gramos y su calibre fue de 36.5 unidades, entonces el sobrepeso mix sería del 5.64%.

6. 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 de indicadores como averías, reproceso y sobrepeso mix en la Compañía de Galletas Noel; indicadores que día a día se calculan y se presentan a la Dirección de Producción para conocer el estado del proceso y soportar la toma de decisiones. Estos tres indicadores son los de mayor importancia, ya que en el caso del reproceso brinda información sobre la cantidad de galleta no conforme, lo que en términos económicos afecta directamente a la compañía, mientras que el sobrepeso mix cuantifica el peso por encima o por debajo de las especificaciones, lo que incurre en costos adicionales de materia prima y material de empaque. Por su parte, las averías constituyen un aspecto importante en términos de programación de la producción y pérdidas por paros en los equipos.

De los resultados se encontró que el turno de trabajo no influye de forma significativa en la probabilidad de ocurrencia de una avería mecánica, según el modelo de regresión logística binaria ajustado. Para trabajos futuros podría tomarse como base dicho modelo para encontrar probabilidades de ocurrencia de averías en equipos más específicos y teniendo en cuenta más variables independientes.

Por otro lado, se encontró que el modelo de bosques aleatorios tiene una mejor eficiencia para predecir los kilogramos de reproceso en el área de multiempaque, y las variables con mayor importancia a la hora de explicar el recorte son el horno y el tipo de reproceso, siendo los hornos Z02, Z05 y Z12 aquellos en donde se incrementa en mayor medida los kilogramos de reproceso.

Finalmente, el modelo de regresión lineal múltiple fue el que explicó mejor el sobrepeso mix de la galleta Satín Fit taco x 5 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, pero no seleccionado debido al principio de parsimonia. 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.

Como trabajo futuro se podría realizar un análisis más profundo del sobrepeso mix incluyendo variables independientes relacionadas con la calidad y características del trigo, que de forma empírica se sabe que influye en el peso de la galleta, también se espera que el modelo ajustado sirva de base para replicarlo en otras referencias de galletas.

Los resultados aquí obtenidos también fueron presentados mediante una herramienta interactiva de visualización de datos que fue realizada con el paquete `Shiny` de R (Chang et. al., 2018), de manera que fuese lo más entendible posible y que, al mismo tiempo, el usuario pudiera interactuar con los modelos predictivos realizados.

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 en la Compañía de Galletas Noel.

7. REFERENCIAS BIBLIOGRÁFICAS

Akaike, H. (1974). A new look at the statistical model identi_cation. IEEE Transactions on Automatic Control, 19 (6), 716-723.

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

Ariza-López, F. J., Rodríguez-Avi, J., & Alba-Fernández, V. (2018). CONTROL ESTRICTO DE MATRICES DE CONFUSIÓN POR MEDIO DE DISTRIBUCIONES MULTINOMIALES. GeoFocus. Revista Internacional de Ciencia y Tecnología de la Información Geográfica, (21), 215-226.

Barajas, F. H., Naranjo-Dueñas, G., & Monsalve-Lugo, E. (2017). Estimación del rendimiento de orellana mediante modelos Gamlss. Revista de la Facultad de Ciencias, 6(1), 67-82.

Batanero, C., Estepa, A., & Godino, J. D. (1991). *ANÁLISIS EXPLORATORIO DE DATOS: SUS POSIBILIDADES EN LA ENSEÑANZA SECUNDARIA*. Recuperado de https://www.ugr.es/~batanero/pages/ARTICULOS/anaexplora.pdf

Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. Critical care, 9(1), 112.

Big Data Republic. (2017). Machine learning for predictive maintenance: where to start? Recuperado de https://medium.com/bigdatarepublic/machine-learning-for-predictive-maintenance-where-to-start-5f3b7586acfb

Buuren, S. & Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. Statistics in Medicine, 20, 1259-1277.

Cabeza Rodríguez, S. (2013). *Funcionalidad de las materias primas en la elaboración de galletas*. Recuperado de http://riubu.ubu.es/bitstream/10259.1/117/5/Cabeza_Rodriguez.pdf

Compañía de Galletas Noel S.A.S. (2018). Estrategia de la Compañía de Galletas Noel S.A.S. Recuperado de https://www.noel.com.co/lacompania/estrategia

Cornfield, J., Gordon, T. y Smith, W. N. (1961). Quantal response curves for experimentally uncontroled variables. *Bulletin of the International Statistical Institute*, 38, 97-115.

David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2017). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-8. https://CRAN.R-project.org/package=e1071

Franco Nicolás, M., & Vivo Molina, J. M. (2007). Análisis de Curvas Roc: Principios básicos y aplicaciones.

Gartner Tech. (2017). Big Data Analytics - Gartner Tech Definitions. Recuperado de https://www.gartner.com/it-glossary/analytics

Harrington, P. (2012). Machine learning in action. Manning Publications Co.

Hoffman, J. I. E. (2015). Logistic Regression. *Biostatistics for Medical and Biomedical Practitioners*, 601-611.

Kotu, V., & Deshpande, B. (2019). Classification. *Data Science*, 65-163.

LeanSis. (2018). ¿Qué es el OEE? | Mejora Continua. Recuperado de https://www.leansisproductividad.com/que-es-el-oee/

Lee, J., Lapira, E., Bagheri, B., & Kao, H. an. (2013). Recent advances and trends in predictive manufacturing systems in big data environment. *Manufacturing Letters*, *1*(1), 38-41.

Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). caret: Classification and Regression Training. R package version 6.0-84. https://CRAN.R-project.org/package=caret

Metalmecánica Internacional. (2017). 10 pruebas de que el Big Data está

revolucionando la manufactura. Recuperado de http://www.metalmecanica.com/temas/10-pruebas-de-que-el-Big-Data-esta-revolucionando-la-manufactura+119841

Miguel Nhuch. (2017). Transformando con datos la Industria de Manufactura | SG Buzz. Recuperado de https://sg.com.mx/revista/50/transformando-datos-la-industria-manufactura

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). John Wiley & Sons.

N. Cristianini and J. Shawe-Taylor. An introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, 2000.

Prati, R. C., Batista, G. E. A. P. A., & Monard, M. C. (2008). Curvas ROC para avaliação de classificadores. Revista IEEE América Latina, 6(2), 215-222.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rigby, B. & Stasinopoulos, M. (2005). Generalized additive models for location scale and shape. Applied Statistics, 54 (3), 507-554.

Scholkopf B., Burges C. J. C, and Smola A. J., editors, Advances in Kernel Methods-Support Vector Learning, Cambridge, MA, MIT Press, 1999.

Stasinopoulos, M. & Rigby, R. (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R.

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. Journal of diagnostic medical sonography, 6(1), 35-39.

Terry Therneau and Beth Atkinson (2018). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13. https://CRAN.R-project.org/package=rpart

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2018). shiny: Web Application Framework for R. R package version 1.1.0. https://CRAN.R-project.org/package=shiny

Yuan, Y., Ma, G., Cheng, C., Zhou, B., Zhao, H., Zhang, H.-T., & Ding, H. (2018). Artificial Intelligent Diagnosis and Monitoring in Manufacturing. Cornell University.