Diseño de un modelo con uso de machine learning y big data para predecir el desabastecimiento de medicamentos

Andrés Felipe Vásquez Vera

Asesor

Luis Angel Anillo Arrieta

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

En los últimos años se ha explorado el uso de técnicas de machine learning y big data para optimización y automatización de los procesos de abastecimiento de la cadena de abastecimiento farmacéutica, sin embargo, esto no se ha documento extensamente en Colombia para el desabastecimiento de medicamentos, una de las problemáticas más relevantes y con más consecuencias del Sistema General de Seguridad Social en Salud. Para abordar esto se va a predecir el desabastecimiento de medicamentos con el uso de machine learning y big data, para lograrlo se usarán los datos del SISMED de los años 2022 y 2023, enseguida se probaron tres modelos de machine learning (random forest, XGBoost y red neuronal) y se evaluaron con métricas claves (accuracy, F1-score, entre otras) y se construyó un dashboard en Power BI como herramienta prototipo. Dentro de los resultados se encontró que el modelo de random forest obtuve el mejor rendimiento, ya que por ejemplo obtuvo un accuracy de 1 y que la red neuronal tuvo mejor desempeño cuando los datos se agrupaban por trimestre. Igualmente se pudo hacer pruebas con este último algoritmo que no ha sido tan usado como los otros, lo que pone de manifiesto que puede seguir explorando el uso de redes neuronales para el manejo de inventario y la predicción de abastecimiento de medicamentos.

***Palabras clave*:** farmacia, machine learning, red neuronal, desabastecimiento de medicamentos, medicamentos.

Abstract

In recent years, the use of machine learning and big data techniques has been explored for the optimization and automation of supply processes in the pharmaceutical supply chain; however, this has not been extensively documented in Colombia for drug shortages, one of the most relevant and most consequential problems of the General Social Security Health System. To address this, we will predict drug shortages with the use of machine learning and big data, to achieve this we will use SISMED data for the years 2022 and 2023, then we tested three machine learning models (random forest, XGBoost and neural network) and evaluated them with key metrics (accuracy, F1-score, among others) and built a dashboard in Power BI as a prototype tool. Among the results it was found that the random forest model obtained the best performance, since for example it obtained an accuracy of 1 and that the neural network had better performance when the data was grouped by quarter. It was also possible to test this last algorithm, which has not been used as much as the others, which shows that the use of neural networks for inventory management and drug supply prediction can be further explored.

***Keywords*:** pharmacy, machine learning, neural network, drug shortage, drugs.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

El desabastecimiento de medicamentos representa uno de los problemas críticos de los sistemas de salud en el mundo. Cuando un medicamento esencial no está disponible, los pacientes pueden experimentar interrupciones en su tratamiento, lo que compromete su adherencia terapéutica y aumenta el riesgo de reacciones adversas debido a cambios en la medicación. Además, esta situación puede afectar de manera más grave a los países y medianos y bajos ingresos (Martín et al., 2020; Shukar et al., 2021). Por ejemplo, en Colombia en el periodo de 2010 y 2021 se presentaron 219 medicamentos desabastecido, de estos 10 se encontraron dos o más veces desabastecido (Sabogal et al, 2022)

En este proyecto, se plantea usar técnicas de machine learning y big data para predecir el desabastecimiento de medicamentos en el contexto colombiano. Para ello, se utilizarán los datos proporcionados por el Sistema de Información de Medicamentos (SISMED), este fue creado por la Comisión Nacional de Precios de Medicamentos (2006) para monitorizar los precios de compra y venta de los medicamentos en el país y se ha venido usando en los últimos años para evaluar la oferta y demanda de los mismos medicamentos dentro del Sistema General de Seguridad Social en Salud (SGSSS) (Comisión Nacional de Precios de Medicamentos y Dispositivos Médicos, 2023).

A nivel mundial se han encontrado que en países como Canadá se ha trabajado la predicción del desabastecimiento de medicamentos con la evaluación del algoritmo XGBoost (Pall et al, 2023), por su parte en Lituania se evaluó el uso de la regresión dinámica para hacer predicción para suplir las carencias de medicamentos Burinskienė (2019). Por último, en España se diseñó una solución basada en Big Data para guiar la búsqueda dentro de una red de farmacias y encontrar la que tuviera el medicamento disponible y así dispensar oportunamente el

medicamento al paciente (Martín et al, 2020), pero en esta no se muestra el uso de modelos de machine learning.

Justificación

El desabastecimiento de medicamentos representa un desafío significativo para los profesionales de la salud, quienes deben dedicar tiempo a encontrar medicamentos sustitutos, estudiar su relación riesgos/beneficios, y gestionar la disponibilidad de estos medicamentos a través de contratos con los distribuidores (Saedi et al., 2016). Este fenómeno no solo consume recursos valiosos, sino que también pone en riesgo la calidad del cuidado de los pacientes, lo que hace urgente la investigación de nuevas formas de gestionar este problema.

Recientemente, se han propuesto enfoques innovadores, como el uso de big data y machine learning. Liu et al. (2021) utilizaron algoritmos de regresión para identificar las variables clave y su impacto en el desabastecimiento de medicamentos. Sin embargo, estos métodos aún no han sido aplicados en el contexto colombiano, lo que representa una oportunidad para introducir una solución adaptada a las necesidades locales.

Según Nguyen et al. (2021), la aplicación de analítica de datos e inteligencia artificial puede optimizar los costos e inventarios, reducir el desabastecimiento y aportar mayor transparencia a la cadena de suministros farmacéuticos. En este sentido, el presente proyecto busca hacer una contribución significativa al campo de la gestión de medicamentos en Colombia, utilizando datos locales y aplicando un enfoque basado en tecnologías avanzadas como el machine learning. Además, se pretende proporcionar una herramienta útil para la gestión logística de los medicamentos, similar a la propuesta implementada en Canadá por Pall et al. (2023), quienes desarrollaron un modelo basado en machine learning para optimizar el suministro de medicamentos en farmacias.

Se espera que este proyecto proporcione a los gestores farmacéuticos y los servicios de salud en Colombia una herramienta eficaz para gestionar el desabastecimiento y mejorar la logística del suministro de medicamentos, beneficiando así a toda la población.

Objetivos

Objetivo General

Predecir el desabastecimiento de medicamentos en Colombia mediante el desarrollo de un modelo predictivo basado en técnicas de Machine Learning y análisis de Big Data

Objetivos Específicos

Determinar las variables predictoras y objetivos para el modelo de predicción de desabastecimiento de medicamentos.

Encontrar un modelo de machine learning con la suficiente precisión y de fácil implementación para la predicción del desabastecimiento de medicamentos.

Proponer una herramienta para la gestión y toma de decisiones en el abastecimiento de medicamentos para los servicios y gestores farmacéuticos.

Estado del Arte

De acuerdo con (Nguyen, 2021) para disminuir el desabastecimiento de medicamento se han usado algoritmos como arboles de decisión, K- means y Q-learning. Igualmente se ha explorado el uso de otros algoritmos de machine learning como regresión líneal y cuadrática para mejorar la cadena de suministros farmacéutico (Rekabi et al 2023). Lo que muestra el potencial del uso de machine learning para el desabastecimiento de medicamento.

En el contexto colombiano, el SISMED es una herramienta clave que recopila información sobre las ventas y dispensaciones de medicamentos a nivel nacional (Comisión Nacional de Precios de Medicamentos, 2006). A pesar de su importancia para la regulación del mercado, no se han encontrado estudios publicados que utilicen directamente el SISMED para analizar el desabastecimiento. Sin embargo, se ha señalado que los datos recogidos en este pueden ser usados para analizar el comportamiento de oferta y demanda de los medicamentos y así tratar oportunamente el desabastecimiento de medicamentos (Comisión Nacional de Medicamentos y Dispositivos Médicos, 2023)

Este proyecto busca explorar el uso de algoritmos de machine learning para la predicción de desabastecimiento de medicamentos en el SGSSS usando los datos del SISMED, para así contribuir con la exploración de estas herramientas en el contexto colombiano y abrir el camino para su implementación por parte de los servicios y gestores farmacéuticos.

Marco Contextual

El desabastecimiento de medicamentos se viene documentando desde 1920 con el desabastecimiento de insulina en esa época, con una persistencia hasta la actualidad alrededor del mundo, con diferente impacto en países de bajos, medianos y altos ingresos (Shukar et al, 2021).

Dentro de las causas asociadas al desabastecimiento de medicamentos se encuentra la escasez de las materias primas, disminución o interrupción de la manufactura por parte de los fabricantes, aumento de la demanda que sobrepasa la capacidad de producción del medicamento y retiro del mercado por falta de rentabilidad de los productos (Burinskienė, 2019).

En Colombia desde el 2018 el Instituto Nacional de Vigilancia de Medicamentos y Alimentos (Invima) es la entidad encargada de recibir, gestionar y declarar los reportes de desabastecimiento de medicamentos en Colombia (Sabogal et al, 2022).

Dentro de nuevas aproximaciones para mejorar la cadena de suministro farmacéutico se encuentra el uso de modelos de optimización matemáticos y heurísticos unidos con big data análisis (BDA) tal y como hicieron (Goodarzian et al. 2024).

Para (Nguyen,2021) Machine Learning se refiere a algoritmos que adquieren conocimiento relacionado con tareas y buscan un rendimiento basado en los datos de entrada (p.6891). Estos mismos autores señalan que el uso de estos algoritmos puede ser útiles dentro del desabastecimiento de medicamentos y la cadena de suministros farmacéutica. Dentro de la literatura se encuentra trabajos como los de Pall et al (2023) donde se evaluó el uso de XGBoost, el uso de regresión dinámica por parte de Burinskienė, (2019) y el uso de árboles de decisión, máquina de soporte vectorial, entre otros en la cadena de suministro farmacéutica (Havaeji, 2023).

De acuerdo con lo anterior, es importante probar el uso de modelos de machine learning para el desabastecimiento de medicamentos en el SGSSS colombiano, con el fin de seguir este nuevo enfoque y ver su aplicabilidad en nuestro contexto.

Metodología

Conjuntos de Datos Utilizados

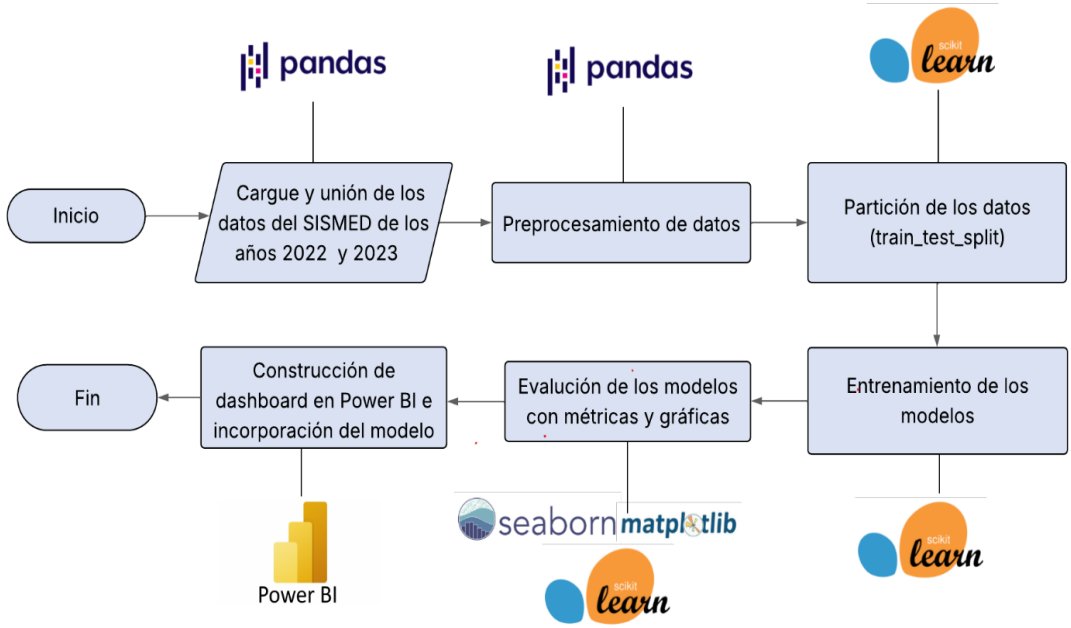
Para el desarrollo de los modelos predictivos se utilizaron los datos reportados de los años 2022 y 2023 del SISMED, lo anterior teniendo en cuenta que son datos abiertos y que representan las transacciones hechas dentro del Sistema General de Seguridad Social en Salud, desde la venta de los medicamentos hecha por fabricantes e importadores hasta las dispensaciones a los pacientes por parte de los gestores y servicios farmacéuticos.

Pasos Generales

En la figura 1 se puede visualizar los pasos generales de la metodología con las librerías de Python usadas en el entorno de Google Colab.

Figura 1

Pasos Generales y Librerías Usadas



Cargue de Datos

Los archivos del SISMED de 2022 y 2023 que se encuentran en archivos de Excel se cargan a colabs y con ayuda de la librería pandas se forma un dataframe, con el fin de facilitar la manipulación de los datos contenidos en el mismo.

Transformación de Datos

Para la transformación de datos se usará pandas con el fin de cambiar tipos de datos, unificar nombres de datos cuando sean iguales, pero tengan diferentes formas de escritura y/o estructura y cambiar nombres de columnas cuando sea necesario.

Conteo de Datos Nulos, en Cero o Vacíos

Se procederá a hacer conteo de celdas vacías, datos nulos o con valor cero, en los que casos en el que caso de que estos se encuentren dentro de las variables predictoras se procederá a eliminar.

Selección de las Variables Predictoras

En el apéndice A se puede encontrar el diccionario de las columnas contenidas en el conjunto de datos del SISMED, a partir de estas se escogerán aquellas que serán las variables predictoras para el modelo de desabastecimiento, teniendo en cuenta los siguientes criterios:

- Columnas relacionadas con las características intrínsecas del medicamento.
- Columnas que den cuenta de las ventas de los medicamentos por parte de los fabricantes e importadores hacía el sistema de salud para su dispensación y uso.
- Columnas relacionadas con las dispensaciones hechas a los pacientes en el SGSSS.
- Columnas que representan unidades de tiempo en términos de mes, año o trimestre.

Adicional a partir de las variables seleccionadas se construirán tres variables, oferta para condensar la cantidad del medicamento vendida en un periodo determinado y disponible para cubrir la demanda, que es la variable que resume la cantidad de dispensaciones realizadas y por último la variable stock acumulado que representa la cantidades acumuladas de un medicamento cuando en un periodo de tiempo sobra o falta medicamento ya sea por disminución o aumento de la demanda y/o disminución de la oferta.

Construcción de la Variable Objetivo

Tomando como base las categorías definidas por el Invima (2023) para el abastecimiento en sus informes periódicos sobre el abastecimiento de medicamentos en Colombia, se construye la variable objetivo estado, en la cual se etiquetaran los datos en dos categorías posibles, una es no desabastecido cuando la oferta más el stock acumulado es mayor o igual a la demanda y se alcanzó a cubrir esta última, por otra parte se clasificará como desabastecido en el caso de que no se haya alcanza a cubrir la demanda.

Codificación y Normalización de Variables

Se usará la función LabelEnconder para la codificación de las variables categóricas y StandarScaler para el escalonamiento y normalización de las variables numéricas.

División y Balanceo de Datos

Se realizará una división en 80% para entrenamiento y 20% de prueba de los datos por medio de la función train_test_splait y en el caso de haber una desproporción en las categorías de la variable objetivo se hará un balanceo en el conjunto de datos de entrenamiento usando la función SMOTE.

Entrenamiento y Prueba de los Modelos Predictivos

Una vez los datos se entrenamiento y prueba se encuentren procesados se procede a hacer el entrenamiento y prueba de los modelos predictivos, para lo cual se usarán los algoritmos de Random Forest, XGBoost y red neuronal por medio la librería Scikit learn.

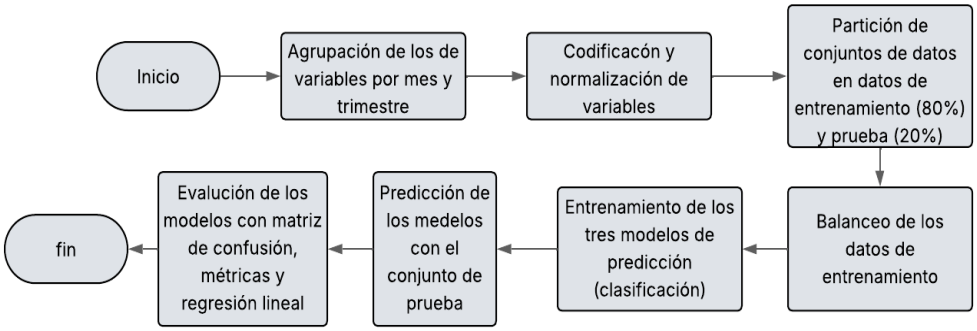
Evaluación y Elección del Modelo Predictivo

Para la evaluación del desempeño de los modelos predictivos, se hará un gráfico de matriz de confusión, el cual se confrontará con el cálculo de las métricas accuracy, presicion, recall y f1-score teniendo en cuenta que estamos usando algoritmos de clasificación. Además, se hará una regresión lineal con el cálculo de R2, error cuadrático medio y raíz del error cuadrático medio con el fin de verificar la relación entre las predicciones hechas por los modelos y la clasificación hecha inicialmente. Todo esto nos llevara a escoger el modelo que presente menos falsos positivos, falsos negativos, métricas más altas y la mejor correlación entre predicción y clasificación real

Los pasos anteriores componen los necesarios para entrenar y evaluar los modelos de predicción del desabastecimiento de medicamentos y se pueden resumir en la figura 2.

Figura 2

Pasos para el Entrenamiento y Evaluación de los Modelos de Predicción



Construcción de Dashboard e Incorporación del Modelo.

Con el dataset que contiene las variables predictoras y objetivo se construirá un dashboard en Power BI y se incorporará al mismo el modelo predictivo escogido.

Resultados

Elección de Variables Predictoras

Del total de las 27 columnas presentes en el conjunto de datos del SISMED de los años 2022 y 2023 se seleccionaron inicialmente 11 columnas las cuales se pueden detallar con sus valores únicos en la tabla 1

Tabla 1

Frecuencia de Posibles Variables Predictoras

Variable	Frecuencia de valores únicos
Periodo	8
Fecha	24
Código tipo de operación	3
Tipo de operación	3
Código tipo de transacción	5
Tipo de transacción	5
Principios activos	13977
Formas farmacéuticas	63
Código unidad factura	4
Unidad factura	4
Total unidades facturadas	145911

Posteriormente se calculó el número de celdas con datos vacíos, nulos o con valor cero presentes en estás once columnas, encontrando que solamente en la columna de principio activo

había 83295 datos nulos; estos fueron eliminados al no representar un principio activo conocido de un medicamento.

Por último, se hizo la agrupación de los datos en dos conjuntos, uno por mes y otro por trimestre, además se agrupo en ambos conjuntos por forma farmacéutica, principio activo, y código de unidad de factura que son las columnas que dan cuenta del medicamento como tal. Para complementar las variables predictoras, se construyó las columnas oferta, demanda y stock acumulado dejando un total de 8 variables predictoras en los dos conjuntos de datos.

Construcción y Etiquetado de la Variable Objetivo

Para la variable objetivo se tomó base las categorías definidas por el Invima (2023) de desabastecido o no desabastecido, para términos de los modelos predictores definimos el estado desabastecido cuando la oferta u oferta más el stock acumulado no suple la demanda y No desabastecido cuando sucede lo contrario.

Enseguida se hizo el etiquetado en estas dos categorías de los datos y se graficó estas proporciones en dos gráficos de barras (figura 3 y 4) para cada conjunto de datos.

Figura 3

Distribución de las Categorías de Abastecimiento de Manera Mensual

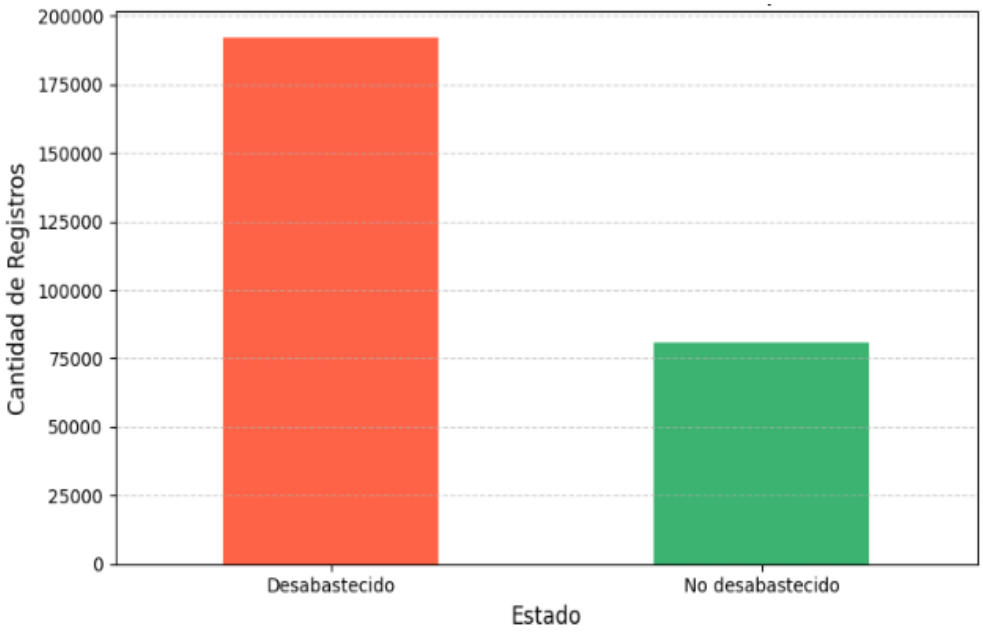
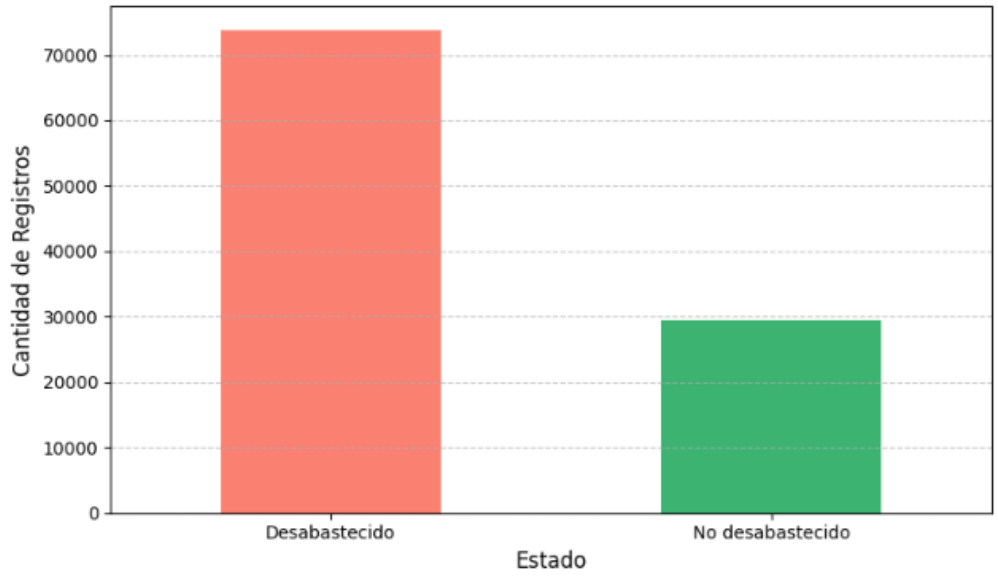


Figura 4

Distribución de las Categorías de Abastecimiento de Manera Trimestral



En la tabla 2 se detallan las variables que se usaran en los modelos predictivos agrupados por cuatro grupos:

Tabla 2

Variables Predictoras y Objetivo

Variables relacionadas con el medicamento	Variables relacionadas con cantidades	Relacionadas con tiempo	Estado de abastecimiento
Principio activo, forma farmacéutica, unidad factura	Oferta, demanda, stock acumulado	Mes, trimestre	No desabastecido, Abastecido

Modelos Predictivos

Para los modelos predictivos se escogieron tres modelos de machine learning random forest, XGBoost y red neuronal simple, como se detalló en la tabla se encontraron 2 medidas de tiempo (mes y trimestre), por lo tanto, las variables se agruparon en dos grupos, uno mensual y otro trimestral. A continuación, se hizo la partición de ambos grupos en conjunto de entrenamiento y prueba, en la tabla 3 se detalla la cantidad de filas que quedaron en cada conjunto de filas que quedaron en los conjuntos de entrenamiento y prueba.

Tabla 3

Distribución de Conjuntos de Entrenamiento y Prueba

Tamaño conjunto de entrenamiento mensual	Tamaño conjunto de prueba mensual	Tamaño conjunto de entrenamiento trimestral	Tamaño conjunto de entrenamiento trimestral
218106	54257	82507	20627

Para los datos de entrenamiento se hizo un balanceo previo dado que como se observa en las figuras 3 y 4 hay un desbalance de clases entre las dos categorías de abastecimiento.

Después se hizo entrenamiento de los tres modelos por duplicado, teniendo en cuenta las agrupaciones por unidad de tiempo mensual y trimestral. Para ver los parámetros que usaron en cada uno de los tres modelos referirse al apéndice B.

Posteriormente se hicieron predicciones con el conjunto de prueba y para la evaluación de estás para cada uno de los algoritmos de predicción de tipo clasificación, se calculó la matriz de confusión para ver la relación entre la clasificación hecha inicialmente y las predicciones y otras cuatro métricas. Las matrices de confusión se muestran en las figuras 5 y 6 y las métricas de evaluación en la tabla 4 para el conjunto de datos mensual y en la tabla 5 para el dataset trimestral.

Figura 5

Matrices de confusión para el Escenario Mensual

Figura 6

Matriz de Confusión para el Escenario Trimestral

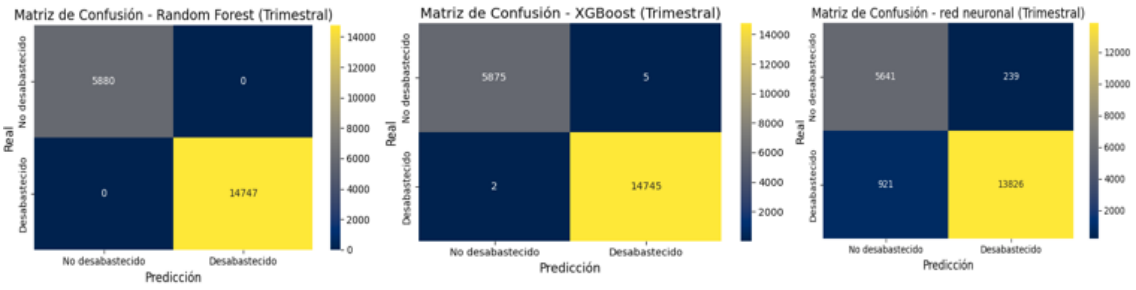


Tabla 4

Métricas de Evaluación de los Modelos en Escenario Mensual

Modelo	Accuracy	Precision	Recall	F1-Score
Random Forest	1.000000	1.000000	1.000000	1.000000
XGBoost	0.999670	0.999792	0.999740	0.999766
Red Neuronal	0.839419	0.830995	0.969004	0.894709

Tabla 5

Métricas de Evaluación de los Modelos en Escenario Trimestral

Modelo	Accuracy	Precision	Recall	F1-Score
Random Forest	1.000000	1.000000	1.000000	1.000000
XGBoost	0.999661	0.999661	0.999864	0.999763
Red Neuronal	0.943763	0.983007	0.937547	0.959739

También se hicieron gráficas de regresión lineal entre la variable estado y las predicciones hechas por cada modelo en los dos agrupamientos (mensual y trimestral) para confirmar la calidad de la relación por medio del cálculo de R-2, error cuadrático medio (ECM) y la raíz cuadrara del error cuadrático medio (RECM). En las figuras 7 y 8.

Figura 7

Gráficos de Regresión Lineal para el Escenario Mensual

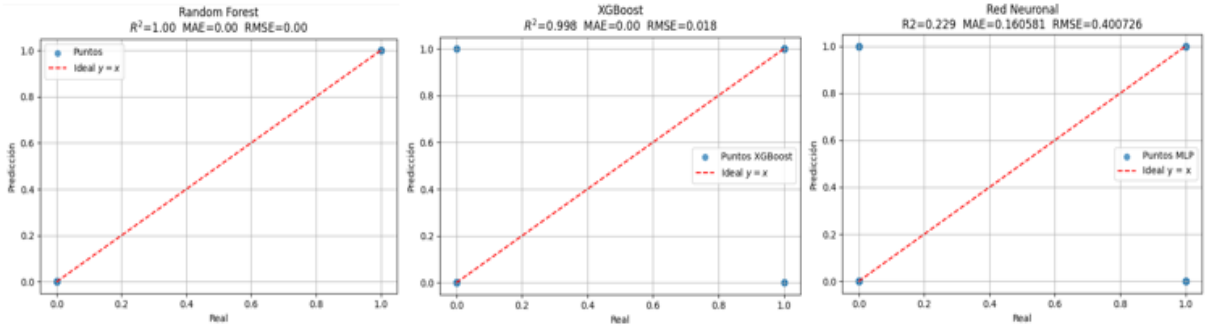
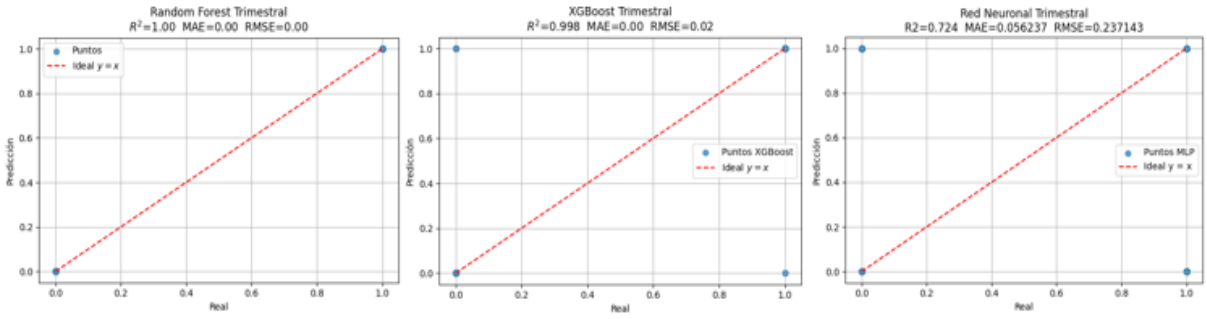


Figura 8

Gráficos de Regresión Lineal para el Escenario Trimestral



Además, en las tablas 6 y 7 se pueden ver las métricas enunciadas anteriormente para las regresiones lineales.

Tabla 6

Métricas de la Regresión Lineal en Escenario Mensual

Modelo	R2	ECM	RCME
Random Forest	1.000000	0.000000	0.000000
XGBoost	0.998416	0.000330	0.018169
Red Neuronal	0.229260	0.160581	0.400726

Tabla 7

Métricas de la Regresión Lineal en Escenario Trimestral

Modelo	R2	ECM	RCME
Random Forest	1.000000	0.000000	0.000000
XGBoost	0.998335	0.000339	0.018422
Red Neuronal	0.724061	0.056237	0.237143

Diseño de un Dashboard en Power BI

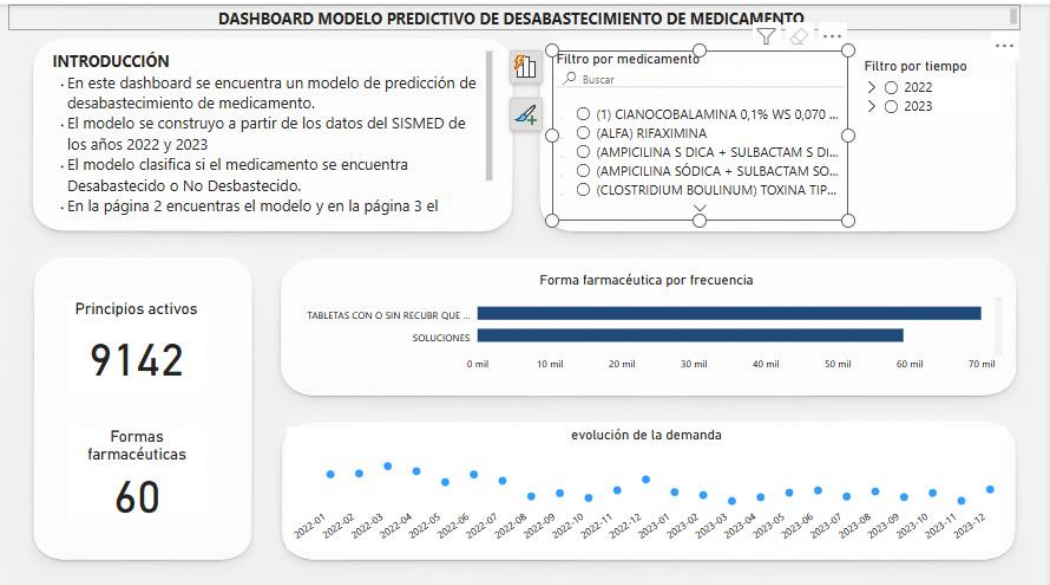
Como herramienta de consulta del modelo predictivo con los resultados más consistente y precisos, el cual fue el Random Forest. Se diseño un dashboard en Power BI de dos páginas, en la primera página se da información general sobre lo datos con los cuales se construyó el modelo y el dashboard, además se puede ver el número total de principios activos y de formas farmacéuticas, un gráfico de barras con la frecuencia de cada forma farmacéutica y uno de dispersión para ver el cambio de la demanda de medicamentos durante los años 2022 y 2023.

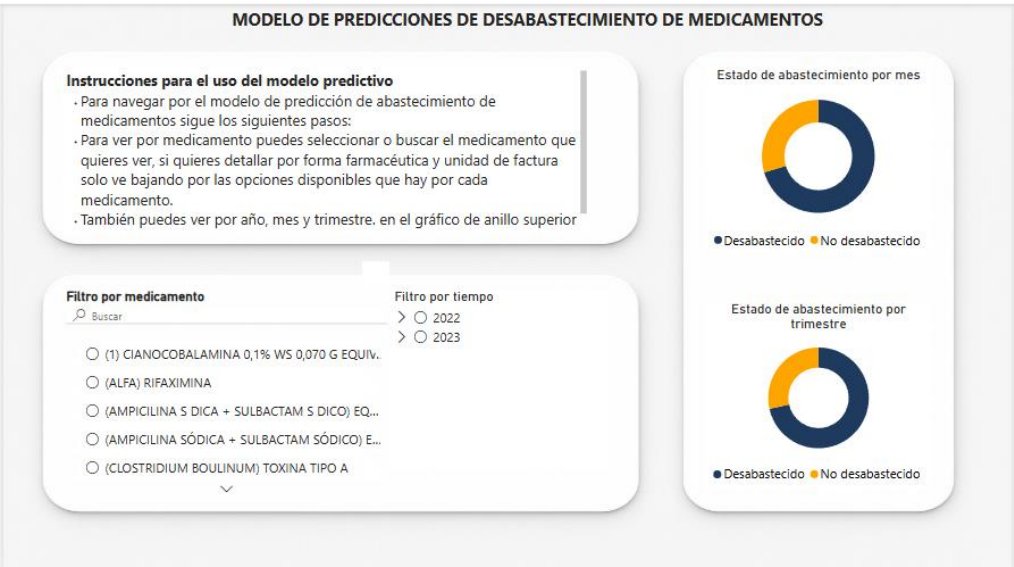
En la segunda página se puede ver la predicción de abastecimiento de los medicamentos en dos gráficos de anillo, uno por mes y el dos para trimestre, con filtros por medicamento y por tiempo.

Las dos páginas del dashboard se pueden visualizar en la figura 9.

Figura 9

Dashboard para el Desabastecimiento de Medicamentos





Discusión de Resultados

Para las variables predictoras encontramos que las columnas principio activo, forma farmacéutica y código de unidad de facturación son las que hablan propiamente del medicamento y por eso se incluyó dentro del grupo de estas variables, sin embargo, se quedan muy limitadas al no tener otros datos como la concentración del medicamento, esto tienen una limitante que consiste en no poder diferenciar por ejemplo para un mismo medicamento diferentes concentraciones y ver la variabilidad de su oferta y demanda, por consiguiente como afectan estás diferentes el estado de abastecimiento.

Lo anterior contrasta con lo hecho por Pall et al. (2023) quienes usaron el Número de Identificación del Medicamento (DNI) otorgado por la agencia reguladora nacional de Canadá para hacer un rastreo de los diferentes movimientos hechos con el medicamento, como lo son compras, dispensaciones y prescripciones, además se usó los movimientos hechos para diferentes medicamentos que son parte de un mismo grupo intercambiable, es decir, medicamentos que se pueden intercambiar sin afectar la efectividad y el tratamiento del paciente. Se podría hacer algo similar a esto agrupando los datos del SISMED por forma farmacéutica, principio activo y concentración para evaluar si se presentan cambios tanto en la variable objetivo al hacer el etiquetado como en el desempeño de los modelos de machine learning.

Referente a las demás variables del modelo se crearon para dar cuenta de los movimientos de compras y dispensaciones y los sobrantes o faltantes de productos en la variable stock acumulado, además de dos variables de tiempo en meses y trimestres, en este último punto se hizo diferente a lo trabajado por Burinskienė (2019) donde se usaron series de tiempo y lags para ver con mayor precisión el abastecimiento de los medicamentos, en contraste como se ven

en las tablas 4 y 5 la red neuronal mejoro su precisión cuando los datos estaban agrupados de manera trimestral.

Para la variable objetivo se hizo una clasificación binaria como no desabastecido y desabastecido que se tomaron de los informes generado por el Invima (2023) sobre el abastecimiento de medicamentos, sin embargo, no se tomaron todas las categorías establecidas por esta entidad debido a que no se cuenta con los datos referentes a la comunicación por parte de importadores o fabricantes de medicamentos cuando comunican si cuentan o no con unidades de estos en un periodo determinado. Además, como vemos en las figuras 3 y 4 al momento de hacer el etiquetado en estos dos estados, para las dos formas de agrupación temporal se forma un desbalance de clases, por lo tanto, se tuvo que hacer un balanceo de los datos de prueba para disminuir el ruido y mejorar la precisión de los modelos de predicción.

En el modelo de Random Forest para ambos escenarios de tiempo mensual y trimestral todas las métricas tienen un valor de 1, esto nos puede decir que el modelo predice muy bien cuando un medicamento esta desabastecido y es real esta alama. Sin embargo, teniendo en cuenta que se hizo un balance en los conjuntos de datos de prueba y que como señalamos anteriormente faltan datos como la concentración que agregan mayor variabilidad al conjunto, se pudo presentar un sobreajuste en este algoritmo que explique estás métricas y, por lo tanto, sería necesario hacer evaluaciones posteriores con datos de otros años para comprobar este mismo desempeño.

En cambio, con el modelo de XGBoost las métricas fueron más bajas, por ejemplo, el acurracy fue de 0.99 y el RCME de 0.18, además como se observan en las figuras 5 y 6 ya se observan tanto falsos positivos como negativos que explican la disminución en las métricas, pero dado que el algoritmo funciona con una función de autocorrección entre los árboles que forma

para corregir errores se puede entender que las métricas hayan bajado, pero si es más factible al usarlo al detector patrones entre los movimientos de oferta y demanda que afecten el abastecimiento de medicamentos y se podría probar con ajustes a parámetros como el número de árboles para ver si mejoran las métricas.

Por último, para la red neuronal se ve una diferente en las métricas entre los escenarios mensual y trimestral, mostrando puntajes más altos en el escenario por trimestre, sobre todo en precisión y recall, lo que quiere decir que en este escenario el algoritmo puede detectar realmente cuando hay un verdadero desabastecimiento de medicamento y la probabilidad de generar un falso positivo es baja. Además, como muestras el R2 y el RECM este modelo mostro mejor desempeño en la agrupación por trimestre, debido a que es más sensible a ruido que se genera en el escenario mensual y que es menor el trimestral.

Adicional es importante señalar que a diferencia de estudios como los hechos por Pall (2023) y Burinskienė (2019) donde se usaron datos correspondientes a dispensaciones y compras hechas en farmacias para paciente ambulatorios, es decir paciente que se les entrega los medicamentos para un tiempo determinado, en nuestro caso en el SISMED se agrupan datos tanto de gestores y servicios farmacéuticos que haces dispensaciones para pacientes ambulatorios y hospitalarios, en este último caso las dispensaciones son por cantidades menores o mayores a las de un ambulatorio y varían de acuerdo a la cantidad de días que reciba el medicamento durante la estancia hospitalaria.

Por lo cual, se puede en futuros estudios hacerlo con datos separados por tipo dispensación y ver cómo se comporta tanto el etiquetado de la variable como los modelos en un escenario hospitalario y en uno ambulatorio por separado.

Conclusiones

Demostramos la factibilidad de hacer predicción con machine learning sobre el desabastecimiento de medicamentos en el contexto del Sistema General de Seguridad Social en Salud colombiano, esto por medio de variables propias del medicamento (principio activo, forma farmacéutica y unidad de factura) y de variables que hablan sobre los movimientos y transacciones de este (oferta, demanda y stock acumulado)

El modelo Random Forest con métricas de 1.0 con los datos de prueba, por lo cual es útil para datos sencillos con baja variabilidad, pero con un riesgo de sobreajuste.

Por su parte el algoritmo XGBoost mostro mayor robustes dado que hace corrección iterativa de errores, manteniendo métricas de acurracy y recall de 0.99.

Por último, la red neuronal obtuvo un recall de 0.96 en el escenario mensual, pero es este mismo escenario se vio afectado por el ruido de los datos, por lo cual para reforzar este se deben trabajar con más datos.

Desarrollamos un prototipo de dashboard en Power BI con alertas tempranas y visualizaciones claras, diseñado para que los servicios y gestores farmacéuticos identifiquen y prioricen rápidamente los riesgos de desabastecimiento.

Recomendaciones

Hacer pruebas piloto en los servicios y gestores farmacéuticos con el fin de hacer implementación los algoritmos de machine learning y así impactar en mejorar de alertas tempranas de pedidos y ahorros de costos por disminución de sobrestock.

Para mejorar el desempeño de los modelos de predicción se debe hacer un identificador compuesto (principio activo, forma farmacéutica y concentración) y hacer un reagrupamiento de las formas farmacéuticas en aquellas que son claves, además de integrar datos sobre flujos de inventario en fabricantes e importadores y separar los escenarios ambulatorios y hospitalarios.

Capacitar a los químicos farmacéuticos en el manejo de herramientas de ciencia de datos y en la interpretación de modelos, de modo que se conviertan en impulsores clave de la integración de machine learning y big data en su práctica profesional.

Referencias Bibliográficas

Burinskienė, A. (2019). use of dynamic regression model for reduction of shortages in drug supply. *Business, Management and Education, 17(2),* 218–231. https://doi.org/10.3846/bme.2019.11297

Comisión Nacional de Precios de Medicamentos. (2006). Circular 04 de 2006 – Decisión de la Comisión Nacional de Precios de Medicamentos. Bogotá D.C.: Comisión Nacional de Precios de Medicamentos.

Comisión Nacional de Precios de Medicamentos y Dispositivos Médicos. (2018). Circular 06 de 2018 - Por la cual se establece el nuevo anexo técnico para realizar el reporte de información al Sistema de Información de Precios de Medicamentos (SISMED) y se dictan otras disposiciones. Bogotá D.C.: Comisión Nacional de Precios de Medicamentos y Dispositivos Médicos.

https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/circular-06-de-2018-cpmdm.pdf

Comisión Nacional de Precios de Medicamentos y Dispositivos Médicos. (2023). Circular 17 de 2023 - Por la cual se modifica el artículo 8 de la Circular 6 de 2018 en relación con el periodo de reporte y plazo para el envío de información al SISMED y se sustituye su Anexo 1. Bogotá D.C.: Comisión Nacional de Precios de Medicamentos y Dispositivos Médicos.

https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/circular-0017-2023-cnpmdm.pdf.

Goodarzian, F., Ghasemi, P., Appolloni, A., Ali, I., & Cárdenas-Barrón, L. E. (2024). Supply chain network design based on Big Data Analytics: heuristic-simulation method in a

pharmaceutical case study. *Production Planning and Control*. https://doi.org/10.1080/09537287.2024.2344729

Havaeji, H., Dao, T. M., & Wong, T. (2023). Supervised Learning by Evolutionary Computation Tuning: An Application to Blockchain-Based Pharmaceutical Supply Chain Cost Model. *Mathematics, 11(9)*, 2021. https://doi.org/10.3390/math11092021

Instituto Nacional de Vigilancia de Medicamentos y Alimentos. (2023). Listado de abastecimiento y desabastecimiento diciembre de 2023. https://www.invima.gov.co/sites/default/files/medicamentos-productos-biologicos/Desabastecimientos/2023/LISTADO%20DE%20ABASTECIMIENTO%20Y%20DESABSTECIMIENTO%20DE%20MEDICAMENTOS%20EN%20SEGUIMIENTO%20-%20DIC%20DE%202023.pdf

Kaakeh, R., Sweet, B. V., Reilly, C., Bush, C., DeLoach, S., Higgins, B., Clark, A. M., & Stevenson, J. (2011). Impact of drug shortages on U.S. health systems. American Journal of Health-System Pharmacy, 68(19), 1811-1819. https://doi.org/10.2146/ajhp110210

Martín Lázaro, R., Castro, L., Molinero, A., & Acosta, J. (2020). Technological solutions of community pharmacies to medicines shortage: Application of the collaborative network model and big data. *Farmaceuticos Comunitarios, 12(4)*, 37–46. https://doi.org/10.33620/FC.2173-9218.(2020/Vol12).004.05

Ministerio de Salud (1995). Decreto 677 de 1996 Por el cual se reglamenta parcialmente el Régimen de Registros y Licencias, el Control de Calidad, así como el Régimen de Vigilancia Sanitaria de Medicamentos, Cosméticos, Preparaciones Farmacéuticas a base de Recursos Naturales, Productos de Aseo, Higiene y Limpieza y otros productos de uso

doméstico y se dictan otras disposiciones sobre la materia. Bogotá D.C.: Ministerio de Salud.
https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=9751

Martín, A., Pérez, J., & Gómez, L. (2020). Impacto del desabastecimiento de medicamentos en la salud pública: Revisión sistemática. Journal of Pharmaceutical Policy and Practice, 13(4), 255-267. https://doi.org/10.xxxx/xxxxxx

Nguyen, A., Lamouri, S., Pellerin, R., Tamayo, S., & Lekens, B. (2022). Data analytics in pharmaceutical supply chains: state of the art, opportunities, and challenges. *International Journal of Production Research, 60(22)*, 6888–6907. https://doi.org/10.1080/00207543.2021.1950937

Pall, R., Gauthier, Y., Auer, S., & Mowaswes, W. (2023). Predicting drug shortages using pharmacy data and machine learning. Health Care Management Science, 26(3), 395–411. https://doi.org/10.1007/s10729-022-09627-y

Rekabi, S., Sazvar, Z., & Goodarzian, F. (2023). A machine learning model with linear and quadratic regression for designing pharmaceutical supply chains with soft time windows and perishable products. *Decision Analytics Journal, 9*. https://doi.org/10.1016/j.dajour.2023.100325

Sabogal De La Pava, M.L., Tucker, E.L. Drug shortages in low- and middle-income countries: Colombia as a case study. J of Pharm Policy and Pract 15, 42 (2022). https://doi.org/10.1186/s40545-022-00439-7

Saedi, S., Erhun Kundakcioglu, O., & Henry, A. C. (2016). Mitigating the impact of drug shortages for a healthcare facility: An inventory management approach. *European*

Journal of Operational Research, 251(1), 107–123. https://doi.org/10.1016/j.ejor.2015.11.017

Shukar, S., Zahoor, F., Hayat, K., Saeed, A., & Gillani, A. H. (2021). Drug shortages: Causes, consequences, and management strategies. *Frontiers in Pharmacology, 12,* 693426. https://doi.org/10.3389/fphar.2021.693426

Tucker, E. L., & Daskin, M. S. (2022). Pharmaceutical supply chain reliability and effects on drug shortages. *Computers and Industrial Engineering, 169*. https://doi.org/10.1016/j.cie.2022.108258

Vásquez, A. [Pharmacistdata]. (25 de mayo de 2025). *Presentación Proyecto de Grado Aplicado Especialización en Ciencia de Datos y Analítica* [Video]. Youtube. https://youtu.be/DC4gjrY4ric

Yadav, S., Singh, S.P. Machine learning-based mathematical model for drugs and equipment resilient supply chain using blockchain. *Ann Oper Res* (2024). https://doi.org/10.1007/s10479-023-05761-0

Yadav, S., & Singh, S. P. (2024). Machine learning-based mathematical model for drugs and equipment resilient supply chain using blockchain. *Annals of Operations Research*. https://doi.org/10.1007/s10479-023-05761-0

Apéndices

Apéndice A

Diccionario de las Columnas Presentes en los Datos del SISMED

Nombre de la columna	Descripción
Periodo	Corresponde a los cuatro trimestres de un año, por ejemplo, el primer trimestre de enero a marzo
Año corte	Corresponde al año reportado, por ejemplo 2022
Mes factura	Se detalla el mes de la factura
Cod rol actor reportante	Es el código del actor que hace el reporte, puede ser 1 o 2
Rol actor reportante	Es la descripción del rol del reportante, es 1 cuando el actor es fabricante o importador de medicamento y 2 cuando es un actor diferente al fabricante e importador del medicamento
Código tipo operación	Es la abreviatura del tipo de operación siendo VN ventas, CM compras y RC recobro
Tipo operación	Corresponde a los tres tipos de operación posibles, ventas, compras o recobros dentro del SGSSS
Código tipo transacción	Es el código del tipo de transacción representada con números de 1 a 5
Tipo transacción	Es la descripción de la transacción, 1 para transacción primaria institucional que son las ventas de medicamentos por parte de fabricantes e importadores y las compras realizadas a estos mismos, 2 son la transacción primaria comercial que son la compra y venta de medicamentos por parte de los fabricantes e importadores y que se hacen con recursos no públicos, 3 son las transacciones secundarias institucionales que son las ventas y compras que se hacen entre actores que no fabricantes e importan con recursos públicos, la 4 son las transacciones secundarias comerciales que son similares a la número 3 pero sin recursos públicos y la 5 es la transacción final institucional que son las ventas realizadas por la dispensaciones hecha a los pacientes con recursos públicos, las compras de estas realizadas por las Empresas Administradoras de Planes de Beneficios (EAPB) y las operaciones de recobro realizadas a la Administradora de los Recursos del Sistema General de Seguridad Social en Salud (ADRES)
IUM 1	Corresponde al primer nivel del Identificador Único de Medicamento (IUM)
IUM 2	Corresponde al segundo nivel del Identificador Único de Medicamento (IUM)
IUM 3	Corresponde al tercero nivel del Identificador Único de Medicamento (IUM)
Número de expediente	Es el campo que hace referencia al número de expediente del medicamento asignado por el Invima
Presentación comercial	Corresponde al consecutivo de la presentación comercial asignado por el Invima
Descripción comercial	Es la descripción de la presentación comercial del medicamento
Forma Farmacéutica	Es la forma física en la cual viene el medicamento y/o el dispositivo que se usa para su administración
CUM código ATC	Corresponde al código alfanúmero de la clasificación anatómica, terapéutica y química (ATC) asociado al código único de medicamento (CUM)
CUM ATC	Corresponde a la descripción del principio activo dada por la clasificación ATC
Principio activo	Corresponde al nombre del compuesto o mezcla de compuesto que tiene la acción farmacológica o terapéutica del medicamento

CUM código vía administración	Son los códigos dados a las vías de administración asociadas al CUM del medicamento
Vía de administración	Corresponde a la descripción de las vías de administración aprobadas por el Imvima para el medicamento
Código unidad factura	Es la letra de la A ha la D para la unidad de en la cual se factura el medicamento
Unidad Factura	Es la descripción de la unidad de factura donde A es presentación comercial, B unidad de embalaje, C es unidad de dispensación y D es unidad mínima de concentración
Precio mínimo unitario	Es el precio mínimo al cual se facturo el medicamento en el mismo mes, tipo de operación y transacción
Precio máximo unitario	Es el precio máximo al cual se facturo el medicamento en el mismo mes, tipo de operación y transacción
Valor total facturado	Corresponde al total facturado para el mes de acuerdo con el tipo de operación realizada
Total unidades facturadas	Corresponde al total de unidades facturadas del medicamento

Nota. Elaboración propia tomando como base lo dispuesto por la CNPM (2018) y el Ministerio de Salud (1995)

Apéndice B

Parámetros de los Modelos Predictivos

Parámetro	Random Forest	XGBoost	Red Neuronal
N estimators	100	100	No aplica
Max depth	No aplica	6	No aplica
Min simples split	No aplica	No aplica	No aplica
Min simples leaf	No aplica	No aplica	No aplica
Learning rate	No aplica	0,1	No aplica
Subsample	No aplica	No aplica	No aplica
Colsample bytree	No aplica	No aplica	No aplica
Use label encoder	No aplica	False	No aplica
Eval metric	No aplica	logloss	No aplica
Hidden layer sizes	No aplica	No aplica	(100,)
Activation	No aplica	No aplica	relu
Solver	No aplica	No aplica	adam
Alpha	No aplica	No aplica	0,0001
Max iter	No aplica	No aplica	300
Random state	42	42	42