PREDICCIÓN DE CUMPLIMIENTO DE ENTREGA DE PEDIDOS FARMACÉUTICOS

Autor(es)

Gustavo Adolfo Montoya Escobar

Tutor

Efrain Alberto Oviedo, Magister (MSc)

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analística y Ciencia de Datos

Medellín, Colombia

2022

Cita	(Montoya,2022)
Referencia **Estilo APA 7 (2020)**	Montoya Escobar, G. (2022). *Predicción de cumplimiento de entrega de pedidos farmacéuticos* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia

Especialización en Analítica y Ciencia de Datos, Cohorte II.

TABLA DE CONTENIDOS

TABLA DE FIGURAS

1. RESUMEN EJECUTIVO

En el desarrollo de esta monografía se trabaja alrededor de un servicio farmacéutico que realiza entregas de medicamentos ordenados a pacientes de diferentes municipios de Antioquia. La gestión logística de estos productos farmacéuticos está condicionada por características particulares de los medicamentos. Algunos medicamentos son enviados bajo una condición especial y estos son recibidos por los pacientes, a través del personal interno de la compañía (domiciliarios), por transportadoras o en un gran porcentaje algunos son recogidos directamente por los pacientes o sus cuidadores.

Para el desarrollo del trabajo se generó un registro de información de las entregas de pedidos de los últimos tres meses, para pacientes con diferentes patologías. Gran parte del éxito en el tratamiento de los pacientes está relacionado con la adecuada adherencia de estos a las recetas generadas posterior a la consulta con el médico prescriptor, sin embargo, esta también depende de que los medicamentos sean entregados de forma oportuna y a tiempo, siendo esta una de las principales causales que los pacientes manifiestan como justificación para la no toma de los medicamentos, pues se ven afectados si no hay disponibilidad de los fármacos por parte de sus operadores.

El objetivo de esta monografía fue determinar si los medicamentos serán entregados dentro de los tiempos establecidos posterior a la solicitud del pedido, a través de algoritmos de clasificación o si por determinadas condiciones se va a incumplir con la promesa de suministro al paciente.

A través de los modelos Regresión logística, Random Forest, Naive Bayes y Máquina de soporte de vectores se pudo realizar un acercamiento a la analítica de datos, generando diferentes métricas que permitieron encontrar un modelo de predicción para anteponer el cumplimiento de los pedidos.

Con el modelo Random Forest se obtuvo una Área curva ROC de 0.85, facilitando la predicción de cumplimiento y generando como consecuencia que la compañía farmacéutica pueda ser más reactiva para solucionar en el menor tiempo posible los inconvenientes que se pueden presentar.

El Notebook se encuentra en el siguiente repositorio

https://github.com/gmontoya8703/Monografia_Pre

2. DESCRIPCIÓN DEL PROBLEMA

Citando a Haynes y Sackett (2018); la adherencia terapéutica se define como “la medida con la que un paciente modifica su conducta, orientándola hacia la ingesta del medicamento o las medidas recomendadas por el médico”. Con el propósito de determinar qué factores afectan la adherencia, se han realizado múltiples investigaciones, sin embargo, se han identificado más de doscientas variables relacionadas con este factor que dificultan que las prescripciones sean cumplidas por los pacientes.

Una de las variables relacionadas con la poca adherencia a los tratamientos está relacionada con la disponibilidad y la entrega oportuna de los medicamentos. Esto afecta la calidad de vida de los pacientes, generando como consecuencia un control precario de la enfermedad, implicando en la mayoría de los casos un incremento en las complicaciones, repercutiendo no solo en la salud de los pacientes, sino también aumentando las consultas hospitalarias y exámenes médicos adicionales, que se traducen en un aumento en los gastos de las aseguradoras (Dilla, Valladares, Lizán, & Sacristán, 2009).

En una investigación realizada por Ramos Morales, expone que algunos de los factores que afectan la adherencia terapéutica están relacionados con el sistema de asistencia sanitaria, pues en ocasiones cuentan con sistemas de distribución deficientes o con un alto rango de inconvenientes al momento de realizar la entrega de los medicamentos (2015).

Con el objetivo de ilustrar la importancia de la adherencia, se presenta el caso de la hipertensión. Los avances realizados en los últimos años han permitido que se desarrollen

medicamentos con una tasa de efectividad más alta y una técnica de administración más sencilla. A pesar de esos avances, no se ha logrado disminuir la morbilidad cardiovascular debido a la baja adherencia terapéutica, lo que contribuye a la falta de control de la enfermedad. Este caso es similar para una gran gama de enfermedades crónicas, donde se evidencia que mejorar la adherencia generaría un mayor impacto en la salud de la población y reduciría todos los efectos colaterales que el bajo control de la enfermedad genera para el sistema de salud (Ortega Cerda, Sánchez Herrera, Rodríguez Miranda, & Ortega Legaspi, 2018).

El análisis de este estudio pretendió entonces visualizar el nivel de impacto que se genera debido al incumplimiento en las entregas de medicamentos a los pacientes, teniendo en cuenta las principales características de los productos, así como las particularidades puntuales de los actores que hacen parte de la cadena de suministros.

Se realizará la investigación a partir de:

- Recopilación o extracción de la información del ERP de la compañía
- Limpieza de datos
- Análisis de datos
- Algoritmo de clasificación

La empresa podrá realizar una evaluación de nivel de servicio y cumplimiento con las entregas pactadas.

2.1 PROBLEMA DE NEGOCIO

En la literatura médica, existen varias investigaciones que buscan determinar estrategias para mejorar la adherencia de los pacientes a los procesos terapéuticos, en su mayoría estas estrategias van enfocadas en uno de los factores más prevalentes que es el sistema de atención de salud y el paciente, haciendo énfasis en mejorar el servicio prestado para beneficiar al paciente y su proceso (Ortega Cerda, Sánchez Herrera, Rodríguez Miranda, & Ortega Legaspi, 2018).

En otra investigación, se conocen algunos procedimientos donde se ha buscado entender los errores en las órdenes de medicamentos a través del análisis manual de ciertos ejemplos de errores comunes, sin embargo, se ha determinado que esta metodología es ineficiente porque no incluyen el rango completo de factores que pueden afectar el cumplimiento de entrega, considerando que los proveedores de medicamentos utilizan sistemas computarizados que tienen intrínsecos factores de riesgos. El desarrollo de este artículo indica que es necesario incluir modelos predictivos que permitan incluir factores que verdaderamente contribuyen a estas fallas que finalmente se reflejan en una baja adherencia terapéutica en los pacientes (King et. al. 2021).

Además, ejemplo de los cambios positivos que puede generar el modelo, se presenta en una investigación realizada en Canadá, donde se utilizó un modelo de Machine learning basado en GANomaly donde dos baselines se entrenaron para aprender sobre órdenes de medicamentos de diez años. Los resultados indicaron que el desempeño del modelo permite convertirlo en una herramienta útil para las farmacéuticas (Hogue et. al. 2021)

Con el objetivo de mejorar la adherencia terapéutica de los pacientes, se convierte en una ventaja competitiva lograr incluir una herramienta que facilite realizar un pronóstico sobre la promesa de servicio ofrecida al cliente. La herramienta permitirá desde el momento en que se reciba la solicitud de medicamentos, determinar si va a ser posible cumplir con los compromisos de entrega de acuerdo con las características de la solicitud o si se van a presentar algún tipo de retrasos en la solicitud. Definir las condiciones de entrega y los tiempos estipulados de la misma representará un gran beneficio para los pacientes, pues disminuirá uno de los principales factores que dificultan la adherencia a los tratamientos.

2.2 APROXIMACIÓN DESDE LA ANALÍTICA DE DATOS

En un mundo cada vez más globalizado donde el activo más importante de las empresas son los datos y el conocimiento que pueden llegar a generar con estos, la compañía busca en la analítica de datos y modelos de machine learning una herramienta que fortalezca sus procesos y genere una ventaja competitiva que seduzca a los clientes, las promesas de una excelente atención y el cumplimento en la dispensación de medicamentos a los pacientes de sus aseguradoras filiales. Para generar una necesidad de entendimiento en los pedidos prescritos; y en obediencia a esta demanda, se quiere utilizar algoritmos de clasificación binaria para realizar predicciones que impacten en la optimización de los procesos de la compañía y permitan una mejora en la eficiencia de entregas de medicamentos a los pacientes.

Los modelos implementados se basan en algoritmos de clasificación, a través de estos se podrá determinar desde el momento de recepción de la solicitud de los medicamentos, si será posible cumplir con la promesa de servicio ofrecida a los pacientes o si existirán algunos contratiempos que impidan que las entregas puedan realizarse en los tiempos esperados.

Con la información generada a través de estos modelos de clasificación, los principales beneficiados serán los pacientes, pues podrán tener un panorama claro sobre la recepción oportuna de los medicamentos y esto permitirá que se aumente la adherencia terapéutica, ya que garantiza el cumplimiento oportuno del tratamiento. Eso también mejorará las condiciones del sistema de salud, pues no habrá efectos colaterales relacionados con el consumo de medicamentos que generan sobre costos al sistema.

2.3 ORIGEN DE LOS DATOS

La información fue suministrada por una empresa farmacéutica, desde una de sus unidades de negocio que en la actualidad se encarga de realizar la entrega de medicamentos prescritos a pacientes de diferentes aseguradoras prepagadas e IPS universitarias. Los datos están totalmente anonimizados con el fin de proteger la información comercial de la empresa e información sensible de los pacientes.

La información corresponde a un rango de tres meses de entregas de pedidos de pacientes de un convenio específico e incluye población que reside solo en el departamento de Antioquia.

2.4 MÉTRICAS DE DESEMPEÑO

El modelo se evaluará bajo las siguientes métricas.

- **Métrica de negocio:** En la actualidad, la empresa farmacéutica no tiene ninguna herramienta de pronóstico que le permita determinar cuál será la promesa de cumplimiento en la entrega de los medicamentos que se puede ofrecer a los usuarios posterior a la solicitud de la orden. Por lo tanto, cualquier porcentaje de predicción será considerado un beneficio para el usuario en su proceso de adherencia terapéutica y para la farmacéutica.

La empresa dispone de unos indicadores construidos con información suministrada por los pacientes donde se evalúa, suministro y adherencia (ver Figura 1).

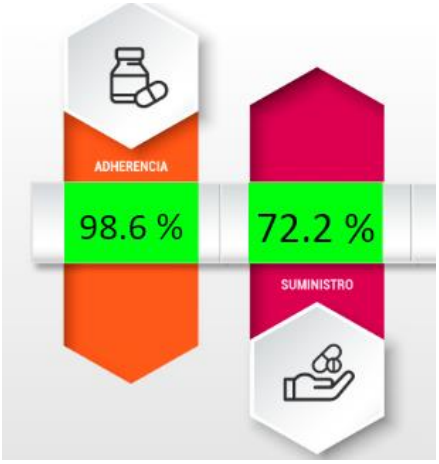


Figura 1. Indicadores generados por los pacientes (creación propia)

- **Suministro:** Porcentaje de pacientes que recibieron el medicamento por parte del asegurador.
- **Adherencia:** Porcentaje de pacientes que reciben el medicamento y este cumple con las condiciones de la prescripción médica.

En la Figura 1 se observa que de toda la población a las que se tenía que entregar las órdenes médicas solo el 72.2 % recibieron el pedido y de estos pacientes que recibieron los medicamentos el 98.6% fueron adherentes al tratamiento. Para entender mejor se puede observar que el 27.8 % de los pacientes que no reciben las órdenes médicas son personas que ya no van a tener una adherencia al tratamiento pues ni siquiera disponen del medicamento.

Con la predicción del cumplimiento de entrega de los pedidos se pretende impactar positivamente en estos indicadores, ya que los procesos de la compañía podrán contar con una

herramienta que permita tomar acciones más preventivas que correctivas en relación de las entregas a los pacientes realizando labores más oportunas y logrando mejorar la disponibilidad del medicamento (suministro) al paciente éste podrá disponer más oportunamente del fármaco y realizar la administración de acuerdo con la prescripción médica, aumentado la adherencia del tratamiento y a su vez mejorando la calidad de vida del paciente.

- **Métrica de machine learning:** Para evaluar los modelos de clasificación se hizo uso de las siguientes métricas de precisión.
    - **Matriz de confusión:** permite visualizar el desempeño de un algoritmo de aprendizaje, facilitando la visualización de los aciertos y errores que se encuentran en el modelo.
    - **Área ROC (Receiver Operating Characteristic):** permite evaluar qué tan bien el modelo puede distinguir dos factores, a través de una gráfica que relaciona la sensibilidad y la especificidad del modelo.
    - **Acurracy:** es una métrica de clasificación que mide el número de predicciones correctas como un porcentaje del total de predicciones realizadas.
    - **F1 Score:** se utiliza para combinar las medidas de precisión y recall en un mismo valor. La presión permite medir la calidad del modelo y el recall o exhaustividad es la cantidad que el modelo puede identificar.

    Como no se dispone de un valor base dentro del negocio se optó por la selección del mejor modelo, el valor de la Área ROC permite ver el comportamiento de las características de las predicciones. Si el valor se encuentra cercano al 1, significa que el modelo está prediciendo correctamente; en cambio, si el valor se encuentra bajo 0.5, se considera que la predicción no es útil.

En este ejercicio se pretende entrenar modelos que generen Áreas ROC por encima de 0.81, cuando los algoritmos de predicción empiecen a mostrar un valor por superior al deseado se empieza a intuir que el rendimiento de los modelos predictivos está siendo óptimo.

3. DATOS

3.1 DATOS ORIGINALES

El conjunto de datos es extraído de la base de datos transaccional de la empresa farmacéutica que dispensa los medicamentos, la información pertenece al periodo de diciembre 2021 a febrero de 2022 donde se encuentran un total de 10445 pedidos, de los cuales en un 68,2% no se cumplió con las condiciones de entrega pactadas.

En la Tabla 1, se presentan las variables originales que contienen los datos utilizados en el modelo, además, se incluye el tipo de variable y la descripción con la información que contiene cada una.

Tabla 1. Datos originales del modelo

Nombre de la Variable	**Tipo**	**Descripción de la variable**
Fecha Pedido	Datetime	Fecha en la que se realizó el pedido por parte del médico prescriptor, IPS o paciente
Dia de la semana	Int64	Día de la semana en número en que se realizó el pedido
Número de Pedido	Int64	Consecutivo del pedido
Entrega Total	Int64	Indica si el pedido se repite en un periodo de tiempo y es un tratamiento de varias entregas
Número Entrega	Int64	Total entregas el pedido a cual corresponde
Porcentaje Tratamiento	float64	Indica el pedido que se está entregando a qué

		porcentaje de tratamiento corresponde
Tipo pedido	object	Prefijo en letras del tipo de pedido
Cantidad Items Pedido	Int64	Cantidad de productos que tienen orden
Origen	object	Medio de recepción o ingreso del pedido
CantItemsCadenafrio	Int64	Productos del pedido que requieren mantener la cadena de frío (los productos en cadena de frío son medicamentos que tienen que ser entregados a los pacientes con unas condiciones especiales, puedes deben conservar ciertas condiciones de temperaturas que implican un empacado específico)
cantAgotados	Int64	Productos que se encuentran agotados (puede haber productos que tengan una alta demanda o que la producción por el laboratorio sea limitada, cuando el laboratorio comunica estas novedades es marcado como agotado, esto no significa necesariamente que el producto no sea entregado al paciente)
cantidad Unidades Pedido	Int64	Suma de todas las unidades de los medicamentos de la orden o pedido
cantidad Unidades Factura	Int64	Suma de todas las unidades que se facturaron
Cantidad de facturas	Int64	Número de facturas que se hicieron para

		completar un pedido.
Cantidad Novedades Entrega	Int64	Número de novedades encontradas al momento de entregar el pedido
Controlado	Int64	Si la dispensación de uno de los medicamentos es controlada por el ministerio de salud o gobierno
Regulado	Int64	Si el precio del medicamento está controlado por el ministerio de salud o el gobierno
Bodega	Int64	Bodega de la cual se está dispensando el pedido
Ciudades	Int64	Ciudad destino de pedido
Departamento	object	Departamento destino del pedido
Distancia Kms	float64	Distancia en kms desde el municipio o residencia del paciente al punto de distribución del pedido
Zona	object	Zona de Antioquia a la que pertenece el paciente
Id Transportador	Int64	Código del transportador
Transportador	object	Nombre del Transportador
Cumple	object	Indica si el Paciente recibe el medicamento en los tiempos pactados (1=Si o 0=No)

3.2 DATASETS

La base de datos origen se encuentra en SQL SERVER y la información se extrae a través de consultas SQL (query) realizando una limpieza de datos inicial de información irrelevante como: pedidos que se entregan en sedes físicas que no pertenecen a la zona de estudio, pedidos anulados y cancelados y datos que son claramente erróneos y que podrían producir ruido. Se revisaron columnas con valores nulos y la información de esta fila se eliminó de la consulta. Se realiza análisis de información no encontrada al momento de hacer uniones (Inner Join y Left Join) con las tablas maestro y las tablas transaccionales, pero no hay inconsistencias entre estas.

No todos los datos que se extraen son necesarios así que algunas columnas son eliminadas directamente desde la consulta y otras se eliminan en los notebooks luego de ver una correlación baja con la variable resultado.

La variable resultado pretende determinar si se cumple o no con la entrega pactada del pedido de los medicamentos. Además, otras variables incluyen datos de características del medicamento (regulados, controlados, cadena de frío, agotados), datos de características del pedido (cantidad de ítems, cantidad unidades pedido, bodega, cantidad de facturas necesarias para completar el pedido), datos geográficos y de transporte (novedades de entregas, municipio de entrega, transportador).

La mayoría de las variables están estructuradas, las variables como número de pedido y origen, que se tuvieron en cuenta en el dataset inicial, son eliminadas ya que sus valores no aportan en la construcción del modelo. El conjunto de datos se divide en datos de entrenamiento y pruebas

con una proporción de 70-30 y al modelo que presenta mejor % en la Curva ROC le aplicaremos la validación cruzada, esta nos ayuda a evitar el sobreajuste del modelo seleccionado.

Como se menciona en el artículo SMOTEMD: Un algoritmo de balanceo de datos mixtos para big data en R, por Morales Oñate, Moreta y Morales Oñate (2020): El análisis de datos desbalanceados se convierte en un reto cuando estos se deben utilizar en términos de modelado. Bajo esta premisa, para la modelación de variables binarias, es necesario usar modelos de probabilidad como logit o probit. Sin embargo, estos modelos pueden ser problemáticos cuando la muestra no está balanceada y se requiere elaborar una matriz de confusión para determinar el poder predictivo del modelo.

3.3 DESCRIPTIVA

Se dispone de un conjunto de datos (CumplimientoPedidos.xlsx) que contiene datos de las entregas de medicamentos a pacientes por parte de un operador durante 3 meses. La compañía se enfrenta a una necesidad evidente de optimizar y mejorar los tiempos de respuesta, ya que el crecimiento en la demanda y el ingreso de nuevos pacientes conlleva a que la logística y la distribución de estos tenga que ser más eficiente.

En un negocio donde el servicio y el tiempo de respuesta son el foco principal de la satisfacción de clientes y usuarios, se encuentra una oportunidad competitiva al aplicar modelos de predicción que ayuden a detectar de manera temprana aquellos vacíos en la operación, y a diseñar y ejecutar acciones correctivas en la cadena de suministro

La variable por predecir es el cumplimento en la entrega, es decir, si el paciente recibió la orden prescrita por su médico en la totalidad (cantidades y número de referencias) (ver Figura

2). En el caso particular de las entregas parciales, estos pedidos serán marcados como incumplidos, porque, aunque en cierta medida podría considerarse que ayudan a mantener la adherencia terapéutica del paciente, este tipo de entregas son muy subjetivas a la patología del paciente. Como ejemplo; las personas con enfermedades mentales como la esquizofrenia deben tener garantizado la disponibilidad y el 100% de tratamiento, así se evidenció en un estudio español que buscaba establecer un consenso sobre cuidados de enfermería para mejorar la adherencia terapéutica en la esquizofrenia (García et al, 2010).

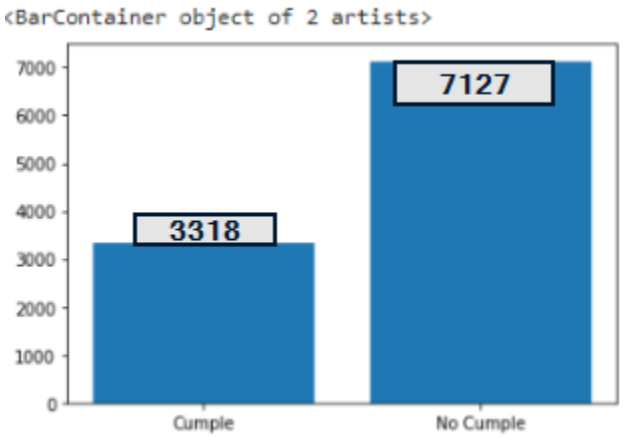


Figura 2. Cantidad de datos que cumplen con las condiciones de entrega.

En la data se registraron 7127 pedidos que no cumplieron con las condiciones de entrega (68.2 %) y 3318 donde se entregaron las órdenes prescritas a los pacientes de forma adecuada (31.8%).

En variables categóricas del dataset como Origen, Zona, Transportador, Convenios, podemos observar que hay datos con mayor peso, ver Figura 3:

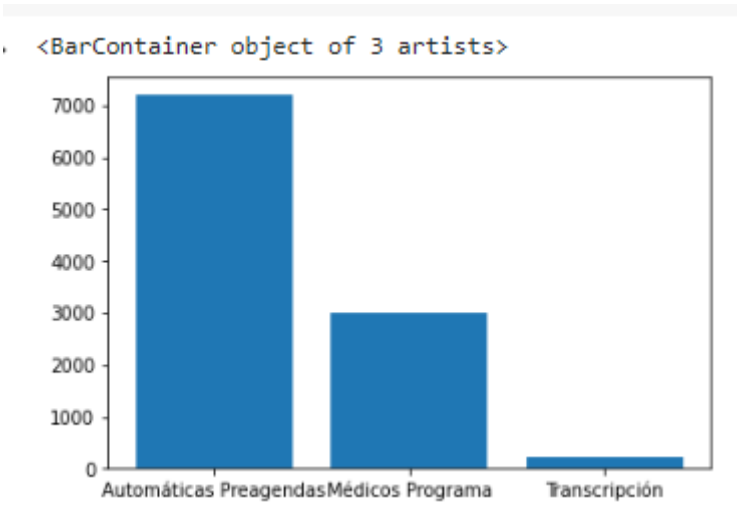


Figura 3. Resumen del origen de los pedidos.

Un gran porcentaje de los pedidos (68.8%), son generados por pre-agendas automáticas.

```
datos_iniciales["Zona"].value_counts()

Valle de Aburrá    9787
Oriente             411
Norte                98
Sin Zona             94
Sur                  31
Occidente            24
Name: Zona, dtype: int64
```

Figura 4.Resumen de la zona donde se deben entregar los pedidos.

En la variable zona se evidencia que gran parte de los pedidos son en el Valle de Aburrá (9787), esta distribución es esperada porque únicamente hay información del departamento de Antioquia y los datos son únicamente de una de las cuentas del operador de medicamentos.

Lo que nos permite simplificar el análisis de los datos, considerando que el operador está ubicado en el municipio de sabaneta (Antioquia) y para este ejercicio se ve que el 93.7% de los pedidos son entregados relativamente cerca.

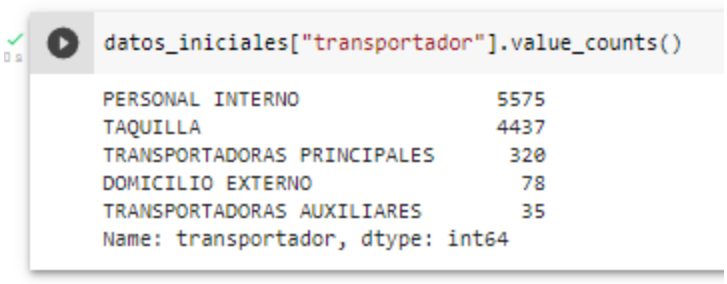


Figura 5. Resumen de los datos del transportador de los pedidos.

Según los datos que se almacenan en la columna transportador (ver Figura 6), se puede inferir que aproximadamente la mitad de los pedidos están siendo entregados por personal de planta del operador y que también un porcentaje importante de pedidos es recogido directamente por el paciente.

4. PROCESO DE ANALÍTICA

Este proceso de analítica se abordó con la finalidad de entender el problema y se realizaron las siguientes fases: preparación, modelación y evaluación de la información (ver Figura 6).

4.1 PIPELINE PRINCIPAL

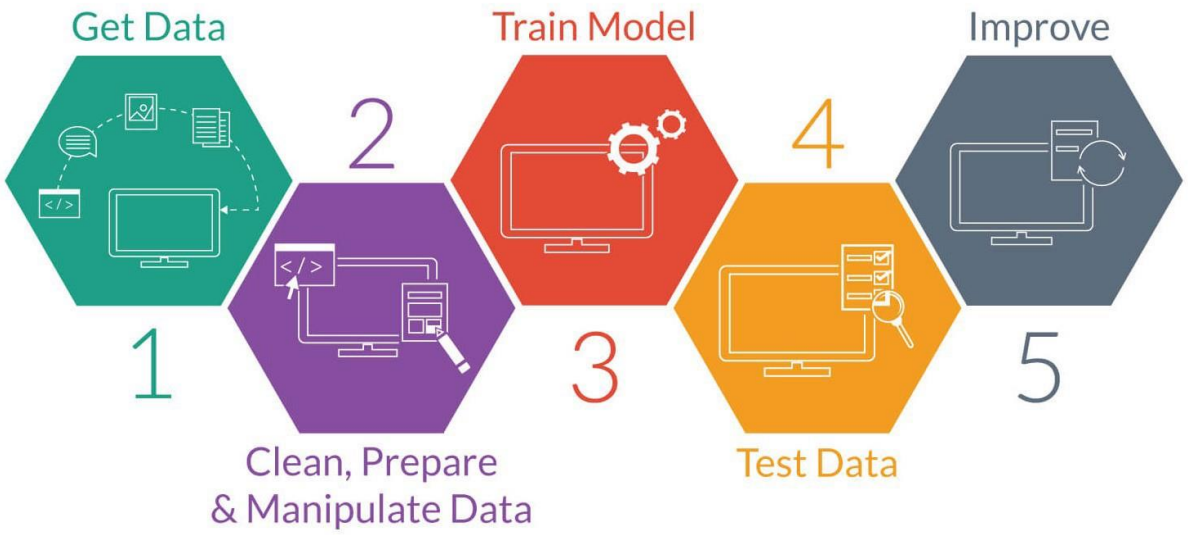


Figura 6. Fases del proceso de analítica (Juana, 2019)

a) **Obtener datos:** previamente a la búsqueda del dataset que se va a implementar, es fundamental que exista claridad respecto a la problemática que se está abordando, además de definir un alcance. En la búsqueda de esta solución, se definen métricas que se utilizan de guía para elaborar los modelos de analítica, siendo estos acordados por los procesos o el cliente que suministra la información. La identificación de los datos, entendimiento de estos y la fuente de donde son obtenidos, son fundamentales para el proyecto de analítica. Se buscó la información de las fuentes, y se realizaron los primeros filtros, eliminando información irrelevante o datos sin calidad que no aportan valor para el análisis.

b) **Limpiar, preparar y manipular los datos:** al obtener los datos en crudo, se realizó un proceso de limpieza, donde se eliminaron los datos incorrectos que podrían generar ruido en el modelo, además se filtraron los datos considerando las variables relevantes para la predicción. Adicionalmente se definieron las características y etiquetas del modelo. En esta etapa se realiza la transformación de variables, la depuración de datos, la conversión de variables y la normalización de datos. Se tuvo en cuenta que:

- La información debía estar anonimizada en caso de que así sea requerido.
- No existieran filas duplicadas e información repetida.
- Hubo homologación de campos.
- Se considerarán campos que funcionan como una descripción.

c) **Entrenar el modelo:** esta etapa se orientó en realizar un análisis de datos, lo cual puede implicar probar diferentes modelos para obtener las mejores métricas.

Se creó un dataset y se dividieron en conjuntos de entrenamiento y prueba. Para esto se utilizaron los modelos de regresión logística, Random Forest, Naive Bayes y máquinas de soporte de vectoreas.

d) **Evaluar o probar el modelo:** posterior al entrenamiento, se evaluó la efectividad del modelo a través de los métodos definidos en la métrica de machine learning.

Uno de los primeros acercamientos para ver qué tan eficiente está siendo el modelo construido es la matriz de confusión.

La matriz de confusión es una herramienta que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase

real, o sea en términos prácticos permite ver qué tipos de aciertos y errores está teniendo el modelo a la hora de pasar por el proceso de aprendizaje con los datos (Arce, 2019).

- **Verdadero positivo:** El valor real es positivo y la prueba predijo también que era positivo.
- **Verdadero negativo:** El valor real es negativo y la prueba predijo también que el resultado era negativo.
- **Falso negativo:** El valor real es positivo, y la prueba predijo que el resultado es negativo.
- **Falso positivo:** El valor real es negativo, y la prueba predijo que el resultado es positivo.

e) **Mejorar el modelo:** a partir de los resultados obtenidos en la evaluación de la predicción se analizan los aspectos a mejorar si se observa que las métricas de evaluación no arrojan resultados eficientes. Se previene que no haya un desbalance de los datos; que la matriz de correlación no haya variables que no aporten al objetivo de los modelos, que los datos de las variables estén normalizados, búsqueda de hiperparámetros de los diferentes modelos y validación cruzada de los mismos.

4.2 PREPROCESAMIENTO

El procesamiento de la información inicia en el momento que se importan los datos, ya sea conectándose a un drive o realizando el cargue directamente desde un archivo de Excel; en este primer acercamiento podemos observar los detalles de los campos del dataset (ver Figura 7), como son:

- Cantidad de información.
- Tipo de campos.

- Campos con datos nulos.

```
[ ]  datos_iniciales.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10445 entries, 0 to 10444
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   numeroPedido              10445 non-null  int64
 1   DiaSemana                 10445 non-null  int64
 2   EntregasTotal             10445 non-null  int64
 3   NumeroEntrega             10445 non-null  int64
 4   PorcentajeTratamiento     10445 non-null  float64
 5   Origen                    10445 non-null  object
 6   cantidadItemsPedido       10445 non-null  int64
 7   CantItemsCadenafrio       10445 non-null  int64
 8   cantAgotados              10445 non-null  int64
 9   cantidadUnidadesPedido    10445 non-null  int64
 10  cantidadUnidadesFactura   10445 non-null  int64
 11  cantidadFacturas          10445 non-null  int64
 12  cantidadNovedadesEntrega  10445 non-null  int64
 13  Controlado                10445 non-null  int64
 14  regulado                  10445 non-null  int64
 15  bodega                    10445 non-null  int64
 16  Distancia Kmts            10445 non-null  float64
 17  Zona                      10445 non-null  object
 18  transportador             10445 non-null  object
 19  convenio                  10445 non-null  object
 20  Cumple                    10445 non-null  object
dtypes: float64(2), int64(14), object(5)
memory usage: 1.7+ MB
```

Figura 7. Detalles del dataset.

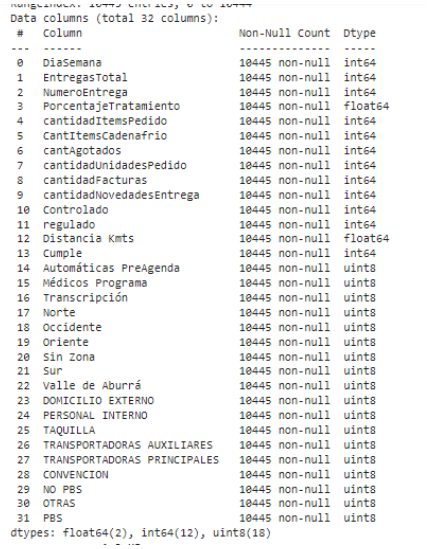
Se eliminaron las columnas número de pedido, tipo, bodega, que no generan valor o que al analizarlas superficialmente se intuya que no va a aportar nada a los modelos de clasificación, en muchos casos estas variables lo que generan es overfitting, su aporte a la predicción se torna irrelevante ya que tanto número de pedido que sus valores es un consecutivo único para cada fila, como el tipo y bodega que sólo difería en dos clases hacen que no sean predictores importantes para los modelos.

Una gran mayoría de modelos de machine learning son más eficientes cuando se ingresan valores numéricos por eso se buscará convertir las variables de texto o categorías; para realizar esta conversión se utilizó onehotencoder, esta idea se refuerza en los análisis de clasificación binaria en conjuntos de datos desbalanceaos realizado en una universidad de Ecuador (Bahamonde y Tapia, 2022)

El método de Label Encoding tiene la ventaja de ser sencillo de implementar. Sin embargo, tiene el problema de que los valores numéricos pueden ser malinterpretados por algunos algoritmos: si hemos codificado varias ciudades con los valores 0, 1, 2 y 3 ¿significa que la ciudad correspondiente al valor 3 es el triple que la que ha recibido el valor 1 (según algún criterio)? La respuesta es no, por supuesto.

Una alternativa al Label Encoding es el método llamado One Hot Encoding. La estrategia que implementa es crear una columna para cada valor distinto que exista en la característica que estamos codificando y, para cada registro, marcar con un 1 la columna a la que pertenezca dicho registro y dejar las demás con 0.

Scikit-Learn implementa esta funcionalidad en la clase sklearn.preprocessing. OneHotEncoder y pandas la implementa en la clase pandas.get_dummies



```
Data columns (total 32 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   DiaSemana                       10445 non-null  int64
 1   EntregasTotal                   10445 non-null  int64
 2   NumeroEntrega                   10445 non-null  int64
 3   PorcentajeTratamiento           10445 non-null  float64
 4   cantidadItemsPedido             10445 non-null  int64
 5   CantItemsCadenafrio             10445 non-null  int64
 6   cantAgotados                    10445 non-null  int64
 7   cantidadUnidadesPedido          10445 non-null  int64
 8   cantidadFacturas                10445 non-null  int64
 9   cantidadNovedadesEntrega        10445 non-null  int64
 10  Controlado                      10445 non-null  int64
 11  regulado                        10445 non-null  int64
 12  Distancia Kmts                  10445 non-null  float64
 13  Cumple                          10445 non-null  int64
 14  Automáticas PreAgenda           10445 non-null  uint8
 15  Médicos Programa                10445 non-null  uint8
 16  Transcripción                   10445 non-null  uint8
 17  Norte                           10445 non-null  uint8
 18  Occidente                       10445 non-null  uint8
 19  Oriente                         10445 non-null  uint8
 20  Sin Zona                        10445 non-null  uint8
 21  Sur                             10445 non-null  uint8
 22  Valle de Aburrá                 10445 non-null  uint8
 23  DOMICILIO EXTERNO               10445 non-null  uint8
 24  PERSONAL INTERNO                10445 non-null  uint8
 25  TAQUILLA                        10445 non-null  uint8
 26  TRANSPORTADORAS AUXILIARES      10445 non-null  uint8
 27  TRANSPORTADORAS PRINCIPALES     10445 non-null  uint8
 28  CONVENCION                      10445 non-null  uint8
 29  NO PBS                          10445 non-null  uint8
 30  OTRAS                           10445 non-null  uint8
 31  PBS                             10445 non-null  uint8
dtypes: float64(2), int64(12), uint8(18)
```

Figura 8. Implementación del one hot encoder.

Después de que la implementación del one hot encoder los valores de las variables categóricas están marcadas con 1 y 0 según corresponda, analicemos la observación del dataset. esto como principio de búsqueda de problemas de desviación.

Se realizó un análisis con la matriz de correlaciones, las variables con valores de coeficiente de correlación más cercanos a 1 muestran una fuerte correlación positiva, los valores más cercanos a -1 muestran una fuerte correlación negativa y los valores más cercanos a 0 muestran una correlación débil o nula.

```
TAQUILLA                      -0.558487
Oriente                       -0.137034
cantidadFacturas              -0.132831
TRANSPORTADORAS PRINCIPALES   -0.117721
Transcripción                 -0.084368
Médicos Programa              -0.077082
PBS                           -0.070915
Sin Zona                      -0.065022
Norte                         -0.062137
cantidadNovedadesEntrega      -0.041605
Sur                           -0.037227
DOMICILIO EXTERNO             -0.035298
NO PBS                        -0.033348
Occidente                     -0.032744
TRANSPORTADORAS AUXILIARES    -0.032447
cantAgotados                  -0.027963
OTRAS                         -0.009443
cantidadItemsPedido           -0.004054
Controlado                     0.003869
cantidadUnidadesPedido         0.007786
PorcentajeTratamiento          0.008154
regulado                       0.016737
EntregasTotal                  0.031485
DiaSemana                      0.033866
CantItemsCadenafrio            0.044801
NumeroEntrega                  0.048393
CONVENCION                     0.095666
Automáticas PreAgenda          0.102522
Valle de Aburrá                0.174379
Distancia Kmts                 0.220429
PERSONAL INTERNO               0.603913
Cumple                         1.000000
Name: Cumple, dtype: float64
```

Figura 9. Correlación con variable objetivo (Si/No cumplimento entrega de pedido prescrito por el médico)

En este ejercicio se evidencia una correlación positiva con la variable objetivo en los campos (Personal Interno, Distancia Kmts, Valle Aburrá, automáticas pre agendas) y correlación negativa con variables como (Taquilla, oriente, Cantidad de facturas, transportadoras principales).

Se utilizaron técnicas de Clustering para analizar cómo se agrupan la información de los pedidos y obtener una variable, que indique un patrón de comportamiento, a pesar de que es una técnica de algoritmos no supervisados podemos implementar algunos de sus modelos para encontrar similitudes y diferencias en los datos, y validar como es el comportamiento de esta variable en los algoritmos de clasificación y si tienen un impacto positivo en los modelos.

El clustering es una técnica para encontrar y clasificar grupos de datos (clusters). Así, los elementos que comparten características semejantes estarán juntos en un mismo grupo, separados de los otros grupos con los que no comparten características.

Para saber si los datos son parecidos o diferentes el algoritmo K-medias utiliza la distancia entre los datos. Las observaciones que se parecen tendrán una menor distancia entre ellas. En general, como medida se utiliza la distancia euclidiana, aunque también se pueden utilizar otras funciones (Duk, 2019).

El K-means es el algoritmo de clusterización que se utiliza en el ejercicio con la finalidad de agrupar los pedidos que conservan características en común, este agrupamiento consiste en reducir la suma de distancias. Con la aplicación de esta técnica surge una nueva variable que da descripción y permite dar entendimiento a dataset.

A través del método del codo como se observa en la Figura 10 nos acercamos a la búsqueda del valor k (número de grupos en que se quiere dividir la información) aunque no hay una forma exacta que determine en cuántos grupos se debe dividir la información, podemos utilizar la técnica del codo y de siluete como se ilustra en la Figura 11 para estimar un número.

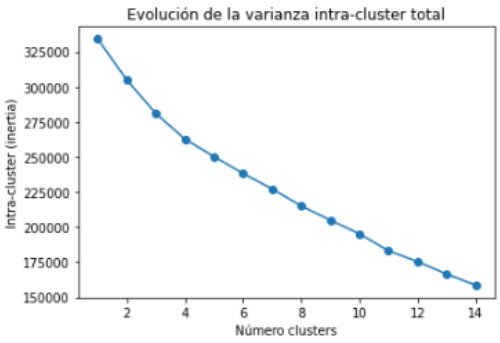


Figura 10. Método del codo

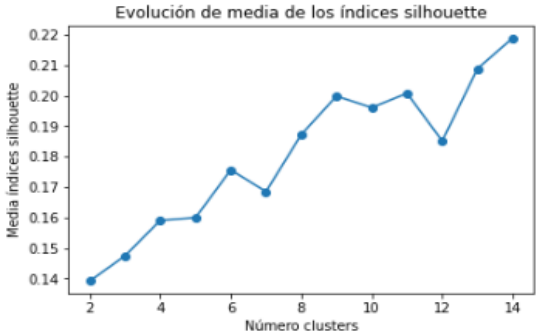


Figura 11. Método del Siluete

A través de los resultados obtenidos en el método del codo y método de siluete se hace una pretensión de agrupar la información en 7 grupos, en la Figura 12 a continuación se puede observar cómo se ha separado la información.

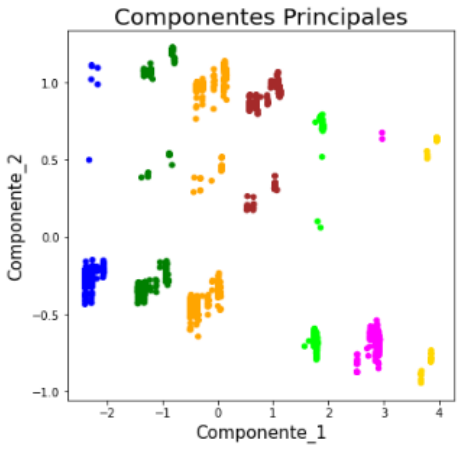


Figura 12. Diagrama de clusterización de los datos

4.3 MODELOS

Este ejercicio se adoptó como un problema de clasificación, se realizó la analítica de los datos con cuatro modelos de predicción, inicialmente en las primeras iteraciones se realizó con los parámetros por defecto y posteriormente en una próxima iteración se hizo la búsqueda de los mejores hiper parámetros.

- **Regresión logística**: Cuando se toma la decisión de trabajar problemas de machine learning de clasificación la regresión logística es una excelente opción, para este ejercicio de predicción de cumplimiento fue seleccionado ya que como método estadístico una de sus fortalezas es medir la predicción de probabilidades binarias.
- **Random Forest:** Basados en los árboles de decisión, el modelo Random Forest es un método predictivo muy estable ya que las métricas no son fuertemente influenciadas por los outliers.
- **Naive Bayes:** El clasificador Naive Bayes asume que el efecto de una característica particular en una clase es independiente de otras características. Por ejemplo, en el momento que se recibe una prescripción médica de un paciente de una de las aseguradoras asociadas, el cumplimiento podrá estimarse de acuerdo a las características de los productos o medicamentos, lugar donde reside, patología, si alguno de los medicamentos solicitados debe conservar la cadena de frío o si este tiene carta de agotados. Incluso si estas características son interdependientes, estas características se consideran de forma independiente. Esta suposición simplifica la computación, y por eso se considera ingenua. Esta suposición se denomina independencia condicional de clase (Gonzalez , 2019).

- **Máquinas de soporte de vectores**: Para Betancourt (2005); "Una Máquina de Soporte Vectorial (SVM) aprende la superficie decisión de dos clases distintas de los puntos de entrada. Como un clasificador de una sola clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con muy poco o ningún conocimiento de los datos fuera de esta frontera. Los datos son mapeados por medio de un kernel Gaussiano u otro tipo de kernel a un espacio de características en un espacio dimensional más alto, donde se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en todas las clases distintas, cada una formando un agrupamiento" (Betancourt, 2005).

4.4 MÉTRICAS

Las métricas utilizadas en este ejercicio, donde el objetivo es dar solución de manera primordial a un tema de suma importancia para una compañía farmacéutica, como lo es la entrega de pedidos oportunamente. Para ello se aborda a través de algoritmos de clasificación supervisada; es importante entender y evaluar que con los datos que va a trabajar la matriz de confusión permite evidenciar el desempeño de los modelos y a través de esta analizar métricas como: Accuracy, la precisión, Recall y la especificidad.

Cuando generamos la matriz de confusión en los modelos encontramos que estos pueden generar errores como falsos positivos que ocurren cuando el algoritmo determina un pedido con la clasificación de que se va cumplir, cuando en realidad no lo hace, y un siguiente error es un falso negativo, cuando el pedido si cumple y todas las condiciones de entregas fueron óptimas y el algoritmo lo clasifica con la etiqueta de incumplimiento.

- Exactitud (Acurracy): Es la suma de predicciones fiables en relación con el total de predicciones. El resultado de este indicador puede estar sujeto a credibilidad si los datos que se ingresan a los modelos no están bien balanceados.
- Precisión: Porcentaje, número o fracción de casos positivos detectados.
- Sensibilidad (Recall): Relación entre las predicciones positivas y correctas y el número total de predicciones.
- Especificidad: Son los casos negativos que fueron clasificados por el modelo de manera correcta.

Para el cálculo de estas métricas se utiliza la librería de Metrics de Sklearn, se importan en el proyecto y la métrica de desempeño se evalúa con el AUC Área ROC.

- from sklearn.metrics import confusion_matrix
- from sklearn.metrics import roc_auc_score
- from sklearn.metrics import accuracy_score
- from sklearn.metrics import f1_score

5. METODOLOGÍA

5.1 BASELINE

La primera y segunda iteración se realizó con dos algoritmos de clasificación (Naive Bayes, Regresión Logística) y a los datos no se les aplicó ninguna técnica de normalización o balanceo, los algoritmos se trabajarán con los hiperparámetros por defectos. El planteamiento de preguntas de ¿cómo elegir los mejores datos para entrenar los modelos? ha sido uno de los primeros cuestionamientos a resolver, pues solo se podrá generar eficiencia en los algoritmos si se trabaja con la información correcta.

Las matrices de correlación es una de las técnicas utilizadas en machine learning para seleccionar las variables correctas, se aclara que la matriz no hace una selección de variables es solo un concepto matemático que se encarga de medir la relación entre estas variables. En el ejercicio como se ve en la Figura 13 a continuación al utilizar esta herramienta encontramos una correlación alta de 0.99 entre las variables cantidad de unidades pedido y cantidad de unidades facturadas.

Figura 13. Matriz de correlación

Se toma la decisión de eliminar de los datos la variable cantidad de unidades facturadas por la gran similitud que poseen sus datos con la variable cantidad de unidades pedidos ya que para que se dé una relación en la información tienen que haber datos diferentes en ambas, sería imposible predecir si el cumplimiento en las entregas de las órdenes médicas está relacionado con la cantidad de unidades del pedido o con la cantidad de unidades facturadas si ambas variables tienen la misma información.

La importancia de los ejercicios realizados en estas primeras iteraciones sirvió como fundamentos para el inicio de una búsqueda de hiperparámetros en los modelos que se desarrollaran, así como observar la necesidad del balance de datos para un correcto entrenamiento.

5.2 VALIDACIÓN

Se podría validar los modelos realizando una repartición de los datos aleatoriamente en dos grupos (Entrenamiento y evaluación), podría ser el medio más práctico y simple, pero esto condiciona la estimación del error ya que su valor es dependiente al conjunto de datos que se utiliza en el entrenamiento y validación.

Los datos ya organizados son divididos en entrenamiento y validación (train-test) 70% - 30% esto se realiza con una función muy práctica de sklearn train_test_split().

La validación cruzada se realizará una vez tengamos seleccionado el mejor modelo, esto permitirá que independientemente de la partición de datos que se haya hecho, sea posible entrenar y validar con todo el conjunto de datos completos. También llamado como cross- validation consiste en dividir los datos de entrenamiento en iteraciones como observamos en la siguiente figura (ver .Figura 14).

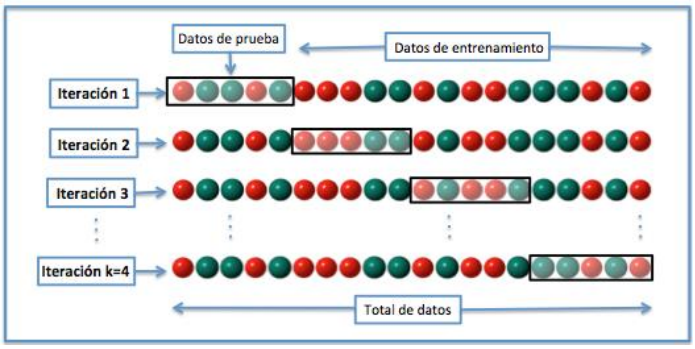


Figura 14. Validación Cruzada, (Fuente: Joan.domenech91 CC BY-SA 3.0)

5.3 ITERACIONES Y EVOLUCIÓN

En las primeras dos iteraciones que se desarrollaron el enfoque estuvo en la limpieza de datos; esta información es extraída de una base de datos transaccional SQL server e importada a un archivo de Excel, se cargan los datos al Notebook y se visualiza la información, se busca la

validación de nulos, de datos incompletos, se borran columnas que no aporten a la analítica, se realiza la matriz de correlaciones.

Se implementaron dos algoritmos el Naive Bayes y la Regresión Logística con su hiperparámetros por defecto los resultados fueron muy bajos estando por debajo de 61.3 %.

En la tercera iteración se realiza la correlación de las variables con la variable objetivo, los resultados muestran coeficientes positivos y negativos, y por consiguiente no se eliminar ninguna variable con base a esta información.

Se busca a través de la técnica de clusterización agrupar la información y obtener una variable que pueda ser implementada con el algoritmo de aprendizaje no supervisado como el k-means. Se normalizan los datos y se utilizan técnicas de balance de datos antes de realizar entrenamiento a los modelos. Cuando en el conjunto de información podemos determinar grupos las relaciones que en la mayoría de veces se hallan, pueden ser muy distintas si las calculamos por cada uno, en este ejercicio no se realizó correlaciones por grupo generado en la clusterización, se hace a nivel general y la mejora obtenida no fue muy significativa.

En la cuarta y quinta iteración, ya con el dataset organizado se trabaja con cuatro modelos Regresión Logística, Random Forest, Naive Bayes y Máquina de soporte de vectores se realiza la búsqueda de los mejores hiperparámetros, se calculan las métricas Accuracy, curva ROC, F1 y se realiza validación cruzada al mejor modelo, en caso de este ejercicio el Random Forest

Cuando aplicamos búsqueda de hiperparámetros se observa una mejora sustancial en el resultado de las predicciones de nuestros modelos.

En este ejercicio se entrena cada modelo, se calculan las métricas Accuracy, F1 score, y curva ROC, comparamos los valores y seleccionamos el mejor.

Una de las dificultades más grandes que se ha tenido es conocer la implicación o el comportamiento que tiene cada hiperparámetro en el modelo.

5.4 HERRAMIENTAS

Para la creación del modelo se utilizó:

- Google Colab para el desarrollo del notebook.
- Librerías de python
  - Pandas
  - Numpy
  - Matplotlib
  - Scikit-learn
- Google Drive
- Github
- Anaconda python
- Jupyter notebook

6. RESULTADOS

6.1 MÉTRICAS

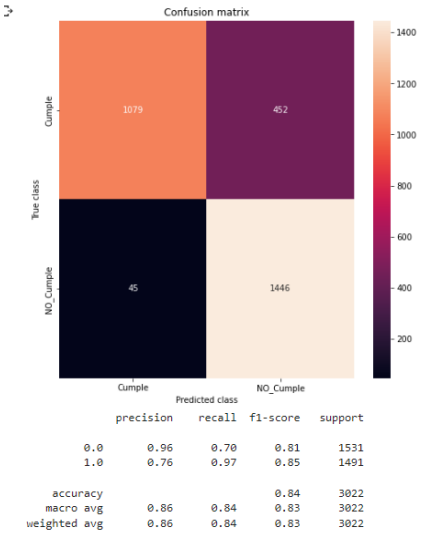
Observemos a continuación la matriz de confusión de cada uno de los modelos utilizados en el ejercicio de clasificación binaria.

- Verdaderos positivos: Cuando el pedido cumple y la predicción dice que el pedido se va cumplir.
- Verdaderos negativos: Cuando el pedido no cumple y la predicción dice que el pedido no se va a cumplir.
- Falsos positivos: Cuando el pedido no cumple y la predicción dice que el pedido se va cumplir
- Falso negativos: Cuando el pedido cumple y la predicción dice que el pedido no se va cumplir

Matriz de Confusión Regresión Logística

- Verdaderos positivos: 1079
- Verdaderos negativos: 1446
- Falsos positivos: 45
- Falso negativos: 452

En la Figura 15 se observa las métricas del modelo que se obtienen con la matriz.



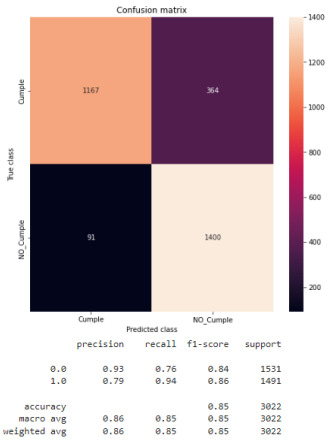
	precision	recall	f1-score	support
0.0	0.96	0.70	0.81	1531
1.0	0.76	0.97	0.85	1491
accuracy			0.84	3022
macro avg	0.86	0.84	0.83	3022
weighted avg	0.86	0.84	0.83	3022

Figura 15. Matriz de Confusión Regresión Logística

Matriz de Confusión Random Forest

- Verdaderos positivos: 1167
- Verdaderos negativos: 1400
- Falsos positivos: 91
- Falso negativos: 364

En la Figura 16 se observa las métricas del modelo que se obtienen con la matriz.



	precision	recall	f1-score	support
0.0	0.93	0.76	0.84	1531
1.0	0.79	0.94	0.86	1491
accuracy			0.85	3022
macro avg	0.86	0.85	0.85	3022
weighted avg	0.86	0.85	0.85	3022

Figura 16. Matriz de Confusión Random Forest

Matriz de Confusión Naive Bayes

- Verdaderos positivos: 1029
- Verdaderos negativos: 1448
- Falsos positivos: 43
- Falso negativos: 502

En la Figura 17 se observa las métricas del modelo que se obtienen con la matriz

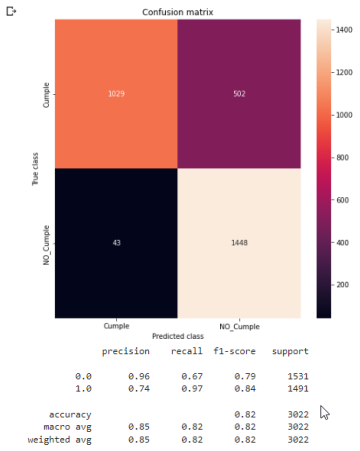


Figura 17. Matriz de Confusión Naive Bayes

Matriz de Confusión Máquina de Soporte de Vectores

- Verdaderos positivos: 1095
- Verdaderos negativos: 1444
- Falsos positivos: 47
- Falso negativos: 436

En la Figura 18 se observa las métricas del modelo que se obtienen con la matriz

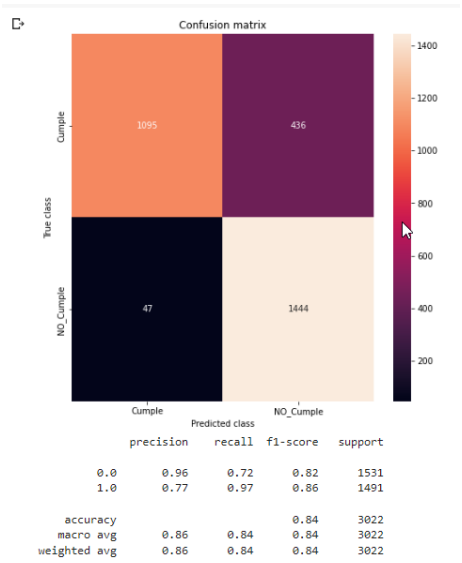


Figura 18. Matriz de Confusión Máquina de soporte de vectores

La tabla comparativa de la métrica de los modelos (ver Tabla 2):

Tabla 2. Comparación de las métricas de los modelos

	Regresión Logística	Random Forest	Naive Bayes	Máquinas de soporte de vectores
Accuracy	0,834	0,857	0,830	0,836
Curva ROC	0,833	0,857	0,830	0,836
F1 Score	0,831	0,856	0,827	0,834

Se selecciona como mejor modelo el Random Forest, dado que representa unos buenos resultados mostrando las mejores métricas y sin problemas de varianza

A el modelo de predicción de pedidos Ramdon Forest seleccionado como el mejor modelo, se le aplica cross- validation con Kernel = Linear, probability = true y seis iteraciones, surgen unos resultados exitosos al ver que las variaciones de las métricas no son muy diferentes, el Accuracy, curva ROC y F1 Score

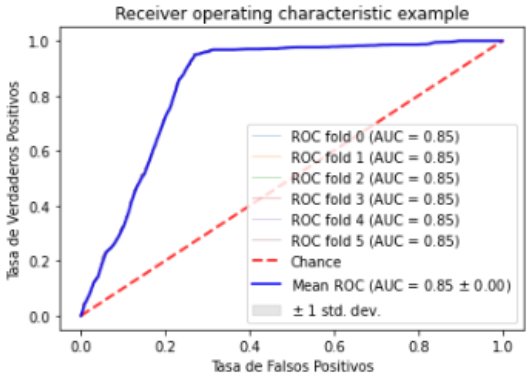


Figura 19. AUC Área bajo la curva ROC - Random Forest

6.2 CONSIDERACIONES DE PRODUCCIÓN

Este ejercicio es 100% académico y una posible implementación no es contemplada por parte de la empresa facilitadora de los datos, argumentando que se quisiera realizar un estudio ampliando el alcance del proyecto con otros convenios asociados.

Sin embargo, se expone que en una posible puesta a producción se debe contemplar servicios en la nube (Azure o Amazon), el modelo podría ser consumido como una API que requeriría una integración con el ERP de la compañía. El modelo necesitaría un reentrenamiento casi que permanente, evaluando constantemente las métricas para evitar sobre entrenamientos; con los diferentes procesos de la compañía se deben realizar un acercamiento a los datos que va evidenciando los algoritmos y a través de la experticia de los equipos tomar decisiones que se puedan contemplar en la selección de información para organizar el modelo y en la reestructuración de procesos de acuerdo con los resultados que se van obteniendo.

7. CONCLUSIONES

En el ejercicio de predicción de cumplimiento de entrega de pedidos farmacéuticos se obtienen resultados muy interesantes a la luz que tiene la compañía como es el mejorar la entrega de medicamentos prescritos a los pacientes y el aporte al aumento de la adherencia.

Para la compañía poder buscar estrategias que permitan encontrar posibles causas de incumplimiento a los pedidos, son herramientas muy ganadoras, ya que las promesas de servicios que se hacen con los clientes es poder ayudarles a que sus pacientes mejoren la calidad de vida y por consiguiente haya una reducción en la requisición de medicamentos, cuando un paciente es adherente a su tratamiento se genera una estabilidad en el control de su patología, evitando así recaer en la enfermedad y por consiguiente se previene a los aseguradores de sobre costos por hospitalizaciones.

A través de los resultados obtenidos, se realiza la sugerencia a la compañía de ampliar la investigación con otros convenios, con otras aseguradoras, además realizar un posible estudio evaluando los datos que fueron anonimizados y posiblemente también teniendo en cuenta otras regiones del país.

Finalmente, para una posible implementación de estos algoritmos de machine learning en la compañía es importante con el equipo de TI y los procesos involucrados; realizar un proyecto serio, con un alcance definido claro, que posiblemente conlleve a reestructuración de procesos y adquisiciones de nueva infraestructura informática, pero es evidente la utilidad que prestaría.

REFERENCIAS

Arce, J. I. B. (2019, julio 26). La matriz de confusión y sus métricas. Juan Barrios. https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

BETANCOURT, G. . A. . (2005). LAS MÁQUINAS DE SOPORTE VECTORIAL (SVMs). Scientia Et Technica, 1(27). https://doi.org/10.22517/23447214.6895

Clement, G. J. (2020, noviembre 30). Naive Bayes algorithms in sklearn. Towards Data Science. https://towardsdatascience.com/why-how-to-use-the-naive-bayes-algorithms-in-a-regulated-industry-with-sklearn-python-code-dbd8304ab2cf

Duk. (2019, enero 8). K-Means Clustering: Agrupamiento con Minería de datos. ESTRATEGIAS DE TRADING. https://estrategiastrading.com/k-means/

Edgar, T. W., & Manz, D. O. (2017). Exploratory Study. En Research Methods for Cyber Security (pp. 95–130). Elsevier.

Gonzalez , L. (20 de Septiembre de 2019). Naive Bayes - Teoría. Obtenido de AprendeIA: https://aprendeia.com/naive-bayes-teoria-machine-learning/

José Alquicira. (2017). Análisis de correlación. (2017, mayo 25). Conogasi. https://conogasi.org/articulos/analisis-de-correlacion-2/

King CR, Abraham J, Fritz BA, Cui Z, Galanter W, Chen Y, et al. (2021) Predicting self-intercepted medication ordering errors using machine learning. PLoS ONE 16(7): e0254358. https://doi.org/10.1371/journal.pone.0254358

Morales Oñate, V., Moreta, L., & Morales Oñate, B. (2020). SMOTEMD: Un algoritmo de balanceo de datos mixtos para Big Data en R. Perfiles, 20-26.

Ortega Cerda, José Juan, Sánchez Herrera, Diana, Rodríguez Miranda, Óscar Adrián, & Ortega Legaspi, Juan Manuel. (2018). Adherencia terapéutica: un problema de atención médica. Acta médica Grupo Ángeles, 16(3), 226-232. Recuperado en 01 de junio de 2022, de

http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1870-72032018000300226&lng=es&tlng=es.

Random Forest con Python by Joaquín Amat Rodrigo, available under a Attribution 4.0 International (CC BY 4.0) at https://www.cienciadedatos.net/documentos/py08_random_forest_python.html

Sophie-Camille Hogue, Flora Chen, Geneviève Brassard, Denis Lebel, Jean-François Bussières, Audrey Durand, Maxime Thibault, Pharmacists' perceptions of a machine learning model for the identification of atypical medication orders, Journal of the American Medical Informatics Association, Volume 28, Issue 8, August 2021, Pages 1712–1718, https://doi.org/10.1093/jamia/ocab071

Sruthi, E. R. (2021, junio 17). Random forest. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Bahamonde Morales, D. I., & Tapia Pisarro, W. S. (2022). Análisis comparativo del rendimiento de algoritmos de clasificación binaria en un conjunto de datos desbalanceados. http://dspace.ups.edu.ec/handle/123456789/22246