UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Clasificación de direcciones usando modelos de Machine Learning: Caso de estudio en una empresa logística ecuatoriana.

Sergio Andrés Recalde Valladares

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
INGENIERO INDUSTRIAL

Quito, 20 de diciembre de 2021

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Clasificación de direcciones usando modelos de Machine Learning: Caso de estudio en una empresa logística ecuatoriana.

Sergio Andrés Recalde Valladares

Nombre del profesor, Título académico María Gabriela Baldeón Calisto, PhD.

Quito, 20 de diciembre de 2021

DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Nombres y apellidos: Sergio Andrés Recalde Valladares

Código: 00131772

Cédula de identidad: 0402129100

Lugar y fecha: Quito, 20 de diciembre de 2021.

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

La industria logística ecuatoriana requiere constantemente de nuevas soluciones e innovación que permita reducir el costo general de la operación. El desarrollo del comercio digital y el reciente florecimiento de la economía ecuatoriana han generado la necesidad de buscar soluciones que se adapten a la creciente demanda y favorezcan la reducción de costos logísticos.

En ese contexto, se estudia el caso de una empresa logística ecuatoriana que busca constantemente la eficiencia operacional, la reducción de costos y aumento de efectividad de entrega. Se propone la automatización de un proceso manual a través de la creación de modelos de Machine Learning que optimicen el uso de recursos como personal y tiempo dedicado. Dicho proceso es un paso inicial en el flujo de entrega de un paquete y funciona a manera de un validador de direcciones para identificar lo que sale inmediatamente a ruta y lo que necesita de información adicional para convertirse en lo que se denomina una entrega efectiva. Así, se utiliza una metodología asociada a la programación de soluciones para identificar claramente el problema, explorar los datos, programar el modelo adecuado y analizar los resultados obtenidos.

Palabras clave: Machine Learning, modelos de clasificación, clasificación de texto, clasificación de direcciones, direcciones.

ABSTRACT

The Ecuadorian logistics industry constantly requires new solutions and innovation that helps reduce the general operational cost. E-commerce boom and recent dynamism of Ecuadorian economy have generated the need of finding new solutions that are adapted to the growing demand and favor logistics costs reduction.
Given this context, a case study is developed in which an Ecuadorian logistics company that is constantly searching for operational efficiency, cost reduction, and effectiveness improvement is set to optimization through code and Machine Learning. In this way, an automatization is proposed in the early stages of the delivery process that will help directly in the use of resources such as workforce and time. The process functions as an addresses validator that separates what is going to be delivered immediately and what is going to be taken on hold until new information is received. A coding methodology is used in order to treat the problem as a development. The first step is to actually understand the problem and then to obtain the data. Once the data is recollected, it needs to be explored and then coded into the best model and finally analyze the results.

Key words: Machine Learning, classifier models, text classifier, address classification, addresses.

Tabla de Contenido

Índice de Tablas

Índice de Figuras

INTRODUCCIÓN

La industria logística a nivel mundial tiene un valor estimado de 8.6 trillones de dólares al año 2020. Siendo el mercado asiático el más grande con un valor de 3098 billones de dólares. La región latinoamericana se encuentra en quinta posición solo por delante de los países CIS (Comunidad de estados independientes). (Mazareanu, 2021).

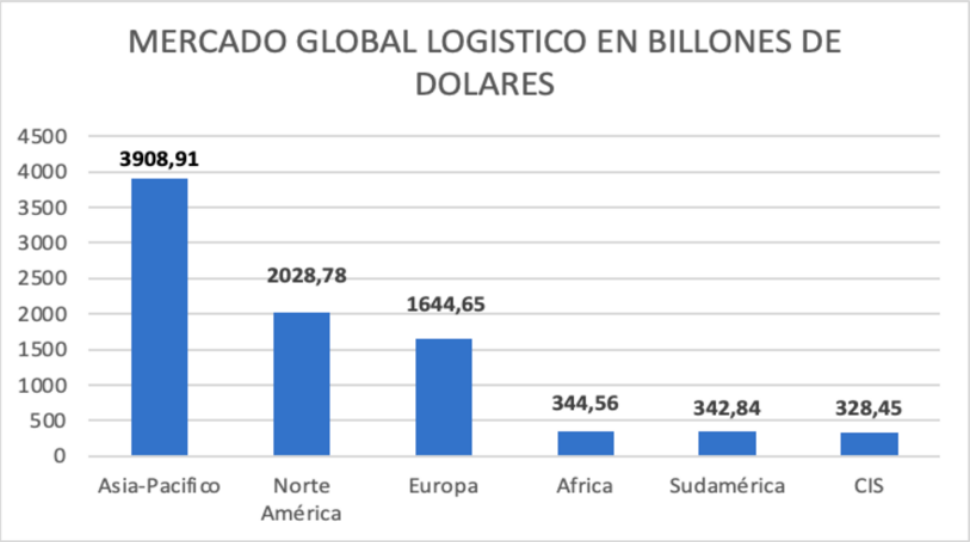

Figura 1: Mercado global logístico en billones de dólares

La industria logística es fuente de grandes innovaciones y desarrollos tecnológicos año tras año debido a que las empresas buscan constantemente abaratar los costos de transporte y optimizar recursos ofreciendo el mismo nivel de servicio a sus clientes. La competitividad del sector también influye directamente en la capacidad de innovar para ganar nuevos clientes y mantener los que se tiene. La infraestructura de transporte y redes logísticas son pieza clave dentro de la conectividad de los diferentes eslabones en la cadena de abastecimiento. En los países de Latino América, las estrategias de competitividad se han basado en mejoras de tarifa, reducción de aranceles y facilitación de procesos. Justamente, la innovación se ve reflejada en la mejora

continua de procesos y optimización de recursos enfocado en el aumento de ventas en clientes y disminución de no conformidades (CEPAL, 2019).

En el Ecuador, la industria logística de servicio de transporte y almacenamiento ha ido creciendo año tras año en el porcentaje de participación dentro del PIB (producto interno bruto) hasta llegar al cuarto puesto solo detrás de los trabajos de construcción, comercialización al por mayor y menor; y de reparación de vehículos y motocicletas y extracción de petróleo, gas natural y actividades de servicio relacionadas. Así, su aportación a la economía ecuatoriana se ha vuelto fundamental y la optimización de recursos de la red logística para ofrecer mejor servicio a los clientes se convertido en prioridad en este tipo de empresas (Banco de Desarrollo de América Latina, 2016). En el año 2016, las empresas logísticas prestaron servicios que correspondían a 410 millones de dólares cuando el mercado demandaba servicios que correspondían a 1238 millones de dólares (Pro-Ecuador, 2016).

Por otro lado, la logística se segmenta en base a al cliente que se atiende, sea el consumidor final (B2C) u otro negocio (B2B). (Tipping & Kauscke, 2016). Así, de la necesidad de innovación y optimización de los procesos internos de las empresas logísticas, nace el objeto de estudio de este trabajo que se enfocará en B2C, específicamente el servicio courier de última milla.

Segmento	Modelo de Negocio	Descripción	Clientes
B2B	LSP	Proveedores de logística	Manufactureras, retailers y centro de distribución
	Carriers	Camiones, carga por tren, por mar y área	LSPs
	CEP	Courier, Servicio Express, Parcel	Retailers, manufactureras y otros clientes
B2C	CEP	Courier, Servicio Express, Parcel	Clientes finales

Figura 2: Segmentación de servicios logísticos

El trabajo se llevará a cabo en una empresa ecuatoriana dedicada a brindar servicios logísticos integrales a clientes tanto nacionales como internacionales. Dicha empresa cuenta con 4 oficinas/hubs de distribución ubicadas en las ciudades principales: Quito, Guayaquil, Cuenca y Ambato. En base a un análisis inicial del proceso operativo de la empresa y entendimiento del negocio en base al seguimiento de la primera fase de la metodología CRISP-DM, se plantea la generación de un algoritmo de Machine Learning para ayuda y automatización de un proceso crítico de clasificación de direcciones de entrega (texto) que se realizaba de manera manual por un operario. En dicho proceso se clasifican direcciones de entrega en base a ciertos parámetros considerados por el encargado del proceso.

DESARROLLO DEL TEMA

Revisión literaria

Se empieza a través de la identificación de Machine Learning en la industria logística para posteriormente recopilar información clave acerca de los distintos tipos de modelos que se pueden utilizar para clasificar texto desde la perspectiva de Machine Learning incluyendo también text classification. Además, se incluye información de Feature Engineering, Ensemble

Learning específicamente el modelo Stacking para combinar varios resultados de distintos modelos de Machine Learning y consolidar un solo resultado más robusto.

Machine Learning en la Logística

Se ha utilizado Machine Learning dentro de la industria logística con el objetivo de automatizar o realizar alguna actividad específica de manera más rápida. Se utiliza especialmente en la generación de rutas óptimas para seguimiento de los courier, también se ha utilizado en manejo de grandes cantidades de datos y obtener información relevante de dicha data. Se utiliza especialmente para realizar Business Analytics y facilitar la toma de decisiones en base a resultados estadísticos. Otro uso que se ha dado a los algoritmos de Machine Learning es en las predicciones de demanda o proyecciones de ventas a realizarse en base a un histórico de datos de años anteriores. A través del uso de Machine Learning se puede tratar de reducir la variabilidad inherente que caracteriza a la demanda y realizar pronósticos mas aterrizados a la realidad. Por otro lado, a través de Machine Learning se pueden recomendar productos para clientes en base a lo que dichos clientes han indicado que les gusta o que aprecian. (Makkar et. al, 2020).

Distintos algoritmos de clasificación

Las técnicas de Machine Learning pueden ser clasificadas en 3 categorías principales:

- **Supervised Learning:** El aprendizaje supervisado es una técnica de Machine Learning en donde se usa información pasada característica de la variable a predecir con el objetivo de enseñar al algoritmo acerca de cuando dicha variable entra en una categoría o la otra. Así, se predice los resultados de una variable en referencia a sus características inherentes. Dentro de los algoritmos específicos que se pueden utilizar para este tipo de aprendizaje se tiene: Naive bayes, Support Vector Machine, regression, Logistic regression, decision trees

y random forests. (Makkar et. al, 2020). Se han utilizado distintos tipos de algoritmos de Supervised Learning para atacar clasificación de imágenes digitales geográficas (Kullkarni, A.), clasificación de tweets y realizar lo que se conoce como sentiment analysis para encontrar el tono de cada mensaje de la red social con ayuda de Naive Bayes (Goel, A. et. al, 2016), o clasificación de imágenes con data de fuentes hídricas utilizando Support Vector Machine (Gidudu, et al, s.f).

- **Unsupervised Learning:** El aprendizaje no supervisado es una técnica de Machine Learning que busca reconocer patrones en los datos de manera autónoma sin participación directa de un operador. Normalmente se divide a la información en grupos segmentados y se genera el análisis de los patrones. Dentro del aprendizaje no supervisado se tiene: Clustering, Reducción Dimensional. (Makkar et. al, 2020).

- **Reinforced Learning:** Este tipo de aprendizaje se basa en procesos reglamentados en donde un algoritmo de Machine Learning es provisto con un conjunto de acciones, parámetros y valores para explorar diferentes opciones y encontrar la mejor. Este tipo de aprendizaje se basa prácticamente en prueba y error hasta que el algoritmo identifica la mejor opción. (Makkar et. al, 2020).

Dentro del alcance de este proyecto se utilizarán métodos de aprendizaje supervisado para identificar la mejor opción entre varios modelos y aplicar dicho modelo a una base de datos de direcciones de entrega que fue compartida previamente por la empresa. Se utilizará este tipo de aprendizaje para aprovechar uno de los campos de la base de datos que actuará como variable target de predicción. Se utilizará también lo que se conoce como Feature Engineering para

descubrir nueva información del campo dirección que ayude a la clasificación y predicción del valor del campo target.

Text Classification

La clasificación de texto utilizando Machine Learning consiste en transformar el texto directamente en tokens y posteriormente en vectores que entran como inputs directos en los modelos ya conocidos de Machine Learning. El transformar un texto en token se conoce como tokenization y es un paso inicial en el proceso posterior de transformar un token a vector (Vectorization). (Luo, 2020). Una vez transformado en vectores, se alimenta un modelo de Machine Learning que realiza la clasificación en base a la data convertida. Sabah et. al en el 2013 mencionan como realizar el proceso de clasificación de texto a través del uso de tokens y simplificación de las palabras (Stemming) dentro de una base de datos de correos electrónicos e identificar si son spam o ham (correos deseados).

Metodología

El presente estudio se enfoca principalmente en la creación de un modelo de Machine Learning que automatice la clasificación de direcciones de entrega de paquetes. Recordando ese contexto, se utilizará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) que es utilizada especialmente para proyectos que requieren el desarrollo de algoritmos o software especializado con enfoque especial en analítica de datos. Dicha metodología se divide en seis pasos específicos que abarcan todo el espectro de un proyecto de esta magnitud, desde el recibir la data hasta manejar adecuadamente los resultados obtenidos. (Azevedo y Santos, 2018). La metodología permite moverse libremente en las distintas fases y reformular ciertos temas de fases anteriores en caso de ser necesario. Es una metodología dinámica que facilita cambios en base a descubrimientos de los datos (Bosjank, 2009).

A continuación, se detallan las fases a seguir en la metodología CRISP-DM.

Fase 1: Entendimiento del negocio

Entender a profundidad el problema en términos del negocio y definir claramente los objetivos que se buscan conseguir de la implementación de la metodología, transformar el entendimiento del negocio en objetivos de obtención de datos específicos sea a través de base de datos compartida o minería de datos. Entender todas las aristas del negocio y la parte específica en la que el proyecto se va a enfocar. (Chapman et al, 2000)

Fase 2: Entendimiento de los datos

Recolectar la información necesaria para poder realizar los modelos de manera posterior. Realizar análisis exploratorios de los datos en donde se pueda identificar las características especiales de la base de datos, temas de correlación entre las variables, identificar si se sigue una distribución específica, categorías, etc. Verificar que los datos que se tienen pueden ser usados de manera óptima en el análisis que se quiere realizar, identificar que los datos no estén sesgados y permitan el desarrollo normal del proyecto. (Azevedo y Santos, 2018).

Fase 3: Preparación de los datos

Seleccionar únicamente la información que realmente va a aportar al estudio por realizar, identificar los datos que no son necesarios y trabajarlos de manera de reducir la variabilidad. Paso previo a la construcción de los modelos en donde se usa los resultados de la fase anterior (Entendimiento de datos) para realizar algún tratamiento especial a los valores nulos, valores repetidos, excepciones, datos atípicos, etc. (Chapman et al, 2000)

Fase 4: Modelado

Utilizar toda la información obtenida de las dos fases anteriores (Entendimiento de los datos y preparación de los datos) para armar los modelos específicos que se considerarían aptos y

que se acoplarían mejor a las características intrínsecas de los datos. Realizar suposiciones en caso de ser necesario y crear los modelos. (Azevedo y Santos, 2018).

Fase 5: Evaluación

Realizar pruebas con los modelos elegidos e identificar las ventajas de un modelo en comparación con otro, averiguar el trade-off entre los distintos modelos y hacer una elección basada en información de relevancia estadística. Identificar si los resultados están alineados al entendimiento del negocio que se realizó inicialmente, si se cumplen los objetivos planteados en las primeras fases de la metodología. Revisar el proceso y determinar los siguientes pasos en términos de los objetivos de negocio. (Chapman et al, 2000)

Fase 6: Despliegue

Realizar el análisis de resultados para identificar mejoras que podrían realizarse en el modelo de clasificación. En caso de que el modelo se encuentre validado y que los datos así lo demuestren, se debería entrar en producción e implementar el modelo. En caso de implementar el modelo, se deben realizar pruebas de que este funcionando acorde a lo estipulado inicialmente. Monitorear el desempeño del modelo de cerca. Realizar un cierre del proyecto identificando los siguientes pasos y documentando toda la información que pueda servir como futura referencia. (Azevedo y Santos, 2018).

EJECUCIÓN DEL TEMA

Fase 1: Entendimiento del negocio

La empresa logística en donde se realizará el proyecto cuenta con el 60% de su operación nacional y el 40% restante con operación internacional. Dicha operación nacional se divide a su vez en 4 servicios especializados: Entregas Certificadas, Servicio in-House, Supply Chain y eFile.

Al hablar de Entregas Certificadas básicamente se hace referencia al servicio de entrega de documentación que requiere el retorno de algún tipo de documento habilitante como contratos, tarifarios, cédulas, planillas de servicio básico, etc. Se puede requerir que algunos de estos documentos habilitantes retornen con la firma de la persona que reciba.

Por otro lado, al hablar de servicio in-House se hace referencia a personal altamente calificado para realizar y manejar procesos internos de clientes y satisfacer requerimientos específicos; servicio de Supply Chain hace referencia a bodegaje en conjunto a picking y Packing; por último, se cuenta con servicios de digitalización de documentos. La mayor cantidad de volumen nacional entra bajo el umbral de Entregas Certificadas. Al analizar el volumen de los cuatro servicios a lo largo del año, se puede identificar que más del 65% del mismo corresponde a Entregas Certificadas. Se puede añadir que los clientes de Entregas Certificadas envían bases de datos masivas para ser entregadas a los beneficiarios.

Es necesario también entender el proceso operativo de la empresa, desde que se reciben las bases de datos hasta que los paquetes salen a ruta para ser entregados. Dentro del proceso descrito existe un paso crítico: identificar de manera manual por un operario que direcciones están con la información completa para poder salir a ruta y ser entregadas por los courier. El encargado de este proceso debe clasificar registro por registro si la dirección se encuentra bien escrita y completa o si la dirección se encuentra mal escrita e incompleta. Lo clasificado como buena dirección se procesa y sale a ruta inmediatamente; lo que sea clasificado como mala dirección no es sacado a ruta y se busca obtener mayor detalle de la dirección a través de procesos internos de la empresa.

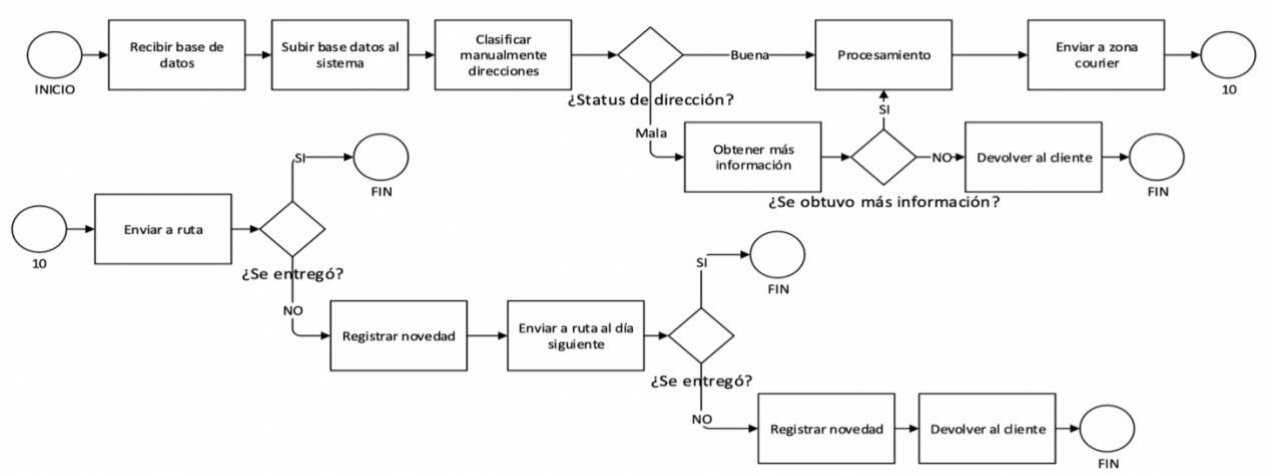

Figura 3: Diagrama de flujo del proceso operativo

Luego de entender el proceso operativo y realizar un análisis de los puntos críticos del mismo, se ha detectado una oportunidad de mejora en la etapa de entrega. Se hace especial énfasis en las actividades previas del proceso, especialmente en la clasificación manual de direcciones (antes de que un paquete salga a ruta). Esta actividad se da cuando se recibe una base de datos por parte de los clientes en donde se tiene toda la información de los beneficiarios de las entregas incluyendo nombres, dirección, teléfonos convencionales y celulares. El siguiente paso es identificar manualmente que dirección es correcta y que dirección es incorrecta en términos de que tan completa se encuentra esa dirección: Calle principal, numeración actual, calle secundaria y referencia, pues en base a esta información el courier realizará la entrega. En caso de que la información este incompleta, se complica la entrega del courier y en muchos casos no se va a poder realizar (regresará a la estación con alguna novedad). Una dirección es correcta si cumple con ciertos parámetros que faciliten la ubicación y entrega del paquete por parte del courier. Así, lo que se clasificó como dirección correcta sale a ruta sin problema con la información inicial compartida por el cliente. Por otro lado, lo clasificado como dirección incorrecta no sale a ruta hasta contar con la información adicional que permita realizar la entrega. Uno de los inconvenientes se da cuando se reciben bases de datos de más de 500 registros y la persona

encargada del proceso debe realizarlo de forma manual, alargando en demasía el proceso. En el caso de que se reciba una base de datos de más 5000 registros la tarea se vuelve aún más complicada de llevar a cabo y en muchos casos ya no se da, creando complicaciones posteriores a los couriers que deben clasificar los envíos antes de salir a ruta, quitándoles tiempo que podría ser usado realizando las entregas y que genera que su ruta se alargue.

Además, la importancia de la clasificación de direcciones radica en la existencia de un costo asociado de las paradas no efectivas de los couriers que incide directamente en el costo de la operación total. Cabe recalcar que no se cobra lo mismo por una parada efectiva que por una parada no efectiva o con novedad. Resulta importante identificar el porcentaje de no entrega sobre todas las entregas certificadas realizadas. Así, al analizar el volumen total de todo el año 2021 se identifica que se ha entregado de manera efectiva el 71% del volumen mientras que el 29% restante no se ha podido entregar.

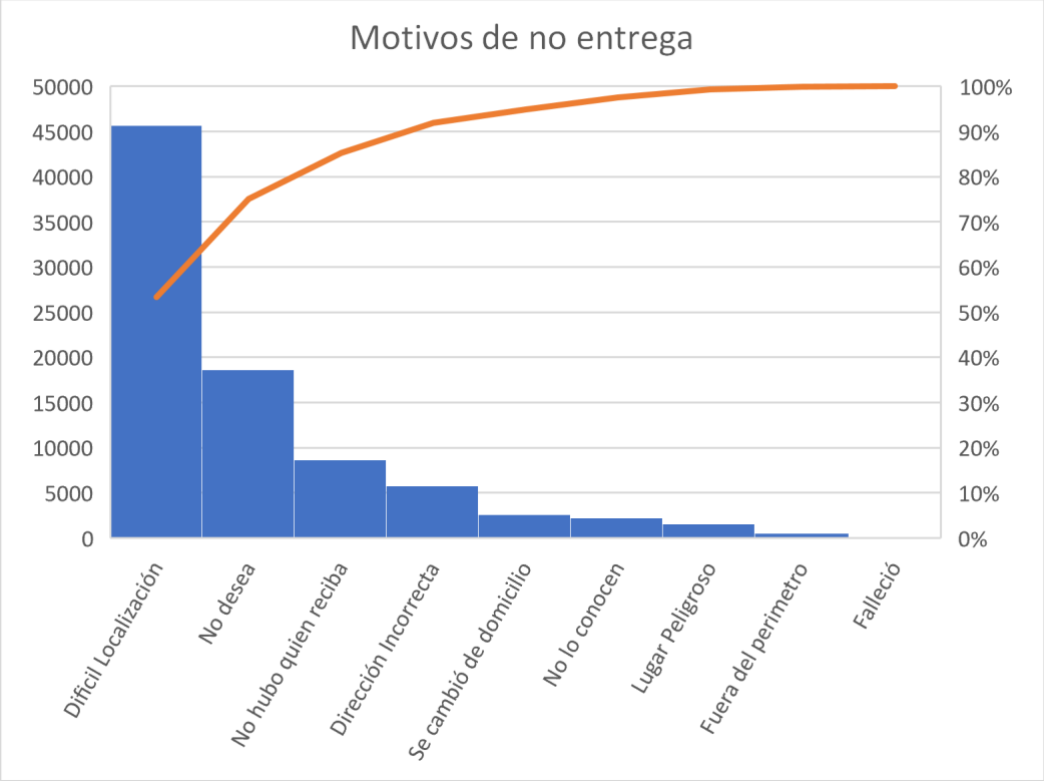

Figura 4: Principales motivos de no entrega

Al desglosar aún más ese 29%, se identifica los motivos de no entrega. Es apreciable que las razones principales son justamente difícil localización que deriva en no encontrar al cliente, no dar con la dirección, no poder realizar la entrega por temas de ubicación. El siguiente motivo es no desea: simplemente el beneficiario final no desea recibir la entrega (motivo no imputable al courier). Otro de los motivos con más porcentaje es no hubo quien reciba, en este caso se llegó a la ubicación del cliente sin embargo no hubo nadie para recibir el envío.

Es importante entender porque el mayor porcentaje de no entregas es la difícil localización. Así, se realiza un análisis de utilizando el diagrama de Ishikawa para determinar las principales causas raíz de estos inconvenientes. Se determina que el envío a ruta de registros con direcciones incompletas o inconformes es una de las principales razones y por detrás de esta razón se tiene la falta de tiempo para clasificar las direcciones cuando llegan bases de datos demasiado extensas lo que genera que esos registros sin verificación pasen a la zona courier de despacho y por ende se realice el intento de entrega sin obtención de información adicional que facilite la ubicación del beneficiario.

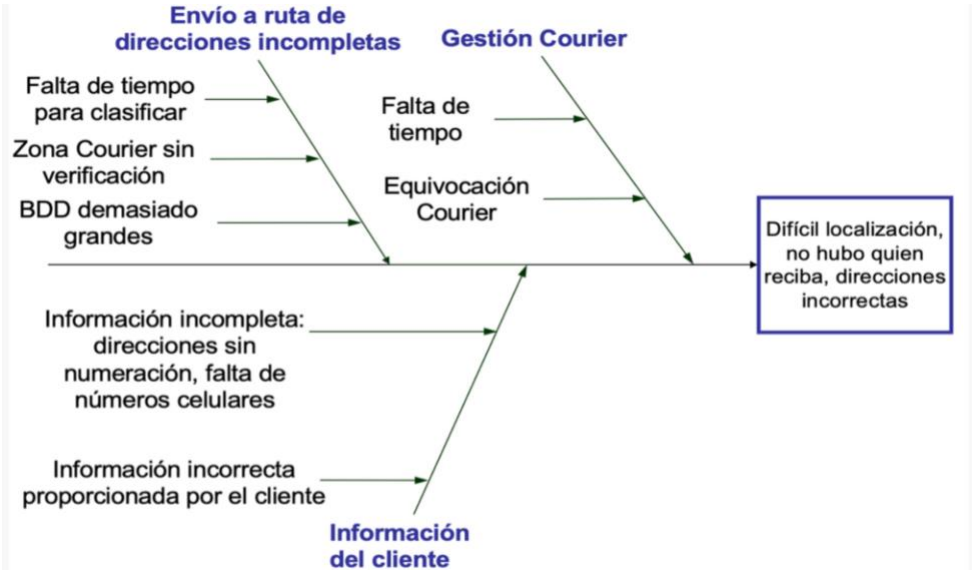

Figura 5: Principales causas raíz de los motivos de no entrega

Una vez entendido como funciona el negocio y los procesos operativos de la empresa, se puede plantear el problema a resolver: Realizar la clasificación de direcciones de manera automática, trabajar en las direcciones incorrectas y evitar que se genere un costo operativo innecesario atado a las entregas no efectivas a través de la construcción de un algoritmo de clasificación de direcciones.

Fase 2: Entendimiento de los datos

Se recibió una base de datos compartida por la empresa con 9347 registros con información de clientes que incluía: nombres, dirección, número de cédula, teléfonos celulares o convencionales e información adicional interna de la empresa como códigos de barra y data de facturación (total 38 campos). Además, se identifica el campo objetivo: Status, en donde se encuentra definido si una dirección es correcta o incorrecta en base a parámetros de clasificación de la persona encargada del proceso manual.

Dentro de los análisis preliminares de la información y para entender mejor la data, se identifica que el 37% (2517 registros) corresponden a Guayaquil, 33% (2222 registros) corresponden a Quito, el 18% (1249 registros) a Cuenca y el restante 12% (789 registros) corresponden a Machala y El Piedrero. Además, se determina que el 65 % de los registros fueron catalogados como dirección correcta y el 35 % restante como dirección incorrecta.

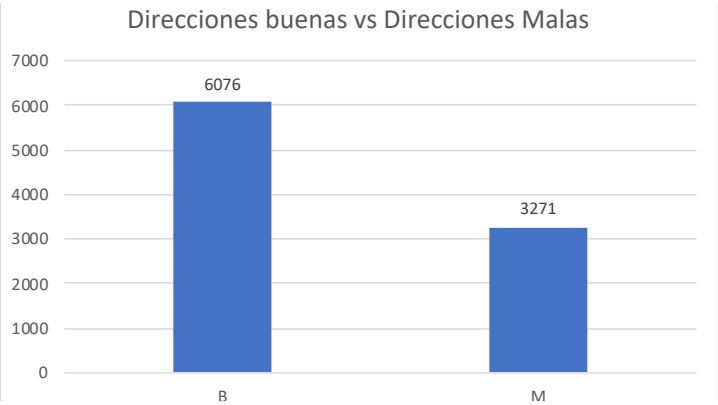

Figura 6: Cantidad de direcciones buenas vs direcciones malas

Al realizar el análisis exploratorio de datos (EDA por sus siglas en inglés) y analizar la correlación entre todas las variables, se determina que ninguna tiene incidencia directa sobre la variable objetivo. Se utiliza un mapa de calor para realizar la respectiva medición. Así, se utiliza Feature Engineering para extraer nueva información útil para la clasificación de la variable a través de la creación de nuevos parámetros (campos en la base de datos). (Zhang y Casari, 2018).

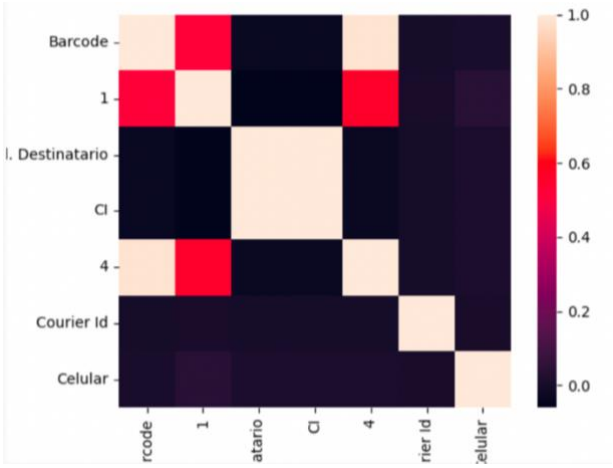

Figura 7: Mapa de calor de variables base de datos

Fase 3: Preparación de los datos

Al identificar que ninguna de las variables se relacionaba directamente con la variable objetivo, se decide identificar los parámetros de clasificación que utiliza la persona que realiza los procesos manuales. Se generan dos sesiones de entendimiento del proceso de clasificación en donde se mapean los parámetros utilizados para catalogar direcciones y se traslada la información aprendida de clasificación a un algoritmo creado específicamente para obtener nueva información a partir de la variable dirección. Se utilizan ciertos parámetros específicos para la ciudad de Quito cómo nomenclatura nueva, referencia, conjunto, urbanización, ciudadela, cooperativa en conjunto con casa, lote, manzana, solar, además de ciertas palabras clave como edificio, bodegas, en el-en la, etc. En el caso de ciudades que no sean Quito, se aplican los mismos parámetros de búsqueda

excluyendo la nomenclatura nueva e intercambiándola por numeración sumando ciertas palabras clave como local, escuela, bomberos, etc.

El algoritmo básicamente realiza un barrido registro por registro siguiendo los parámetros explicados por la persona encargada del proceso y posteriormente letra por letra en cada registro, generando una diferenciación por ciudad y buscando cada una de las palabras mencionadas. En caso de ir encontrando dicha información en cada registro, va asignando un valor categórico en nuevos campos y en caso de no encontrar asigna un valor específico para denotar eso. Así, se reduce la cantidad de campos de la base de datos inicial y se añaden campos que sirven como input a los modelos de Machine Learning puesto que se relacionan directamente al campo dirección y son de utilidad para la clasificación que generen los modelos como output.

```
df.columns
```

```
Index(['Canton', 'Direccion', 'Destinatario ', 'Status', 'Numeracion',
       'ReferenciaCompleta', 'Nomenclatura_Quito', 'Validacion', 'Edificio',
       'En', 'Diagonal', 'Bomberos', 'Hospital', 'Bodega', 'Local', 'Hotel',
       'Complejoturistico'],
      dtype='object')
```

Figura 8: Campos generados con el algoritmo de Feature Engineering

Destinatario	Direccion	Canton	Status	Quito	Provincia	Nomenclatura	Numeracion	Referencia	Urbanizacion
Sergio Recalde	Urbanizacion Jardines de la Pampa lote 179 casa verde de 3 piso	Quito	1	1	0	0	1	1	1

Figura 9: Ejemplo de registro simplificado utilizando el algoritmo de Feature Engineering

Fase 4: Modelado

Una vez obtenida la nueva base de datos reducida dimensionalmente y con campos de información relacionada al campo dirección, se procede con la generación de distintos modelos de Machine Learning. Como se explicó inicialmente, se utilizarán modelos de aprendizaje supervisado debido a que ya se cuenta con el resultado de la variable objetivo. Este resultado fue

generado por la persona encarga del proceso. Dentro de los modelos supervisados, se utilizan los siguientes: Naive Bayes, Neural Networks, SVM, Random Forest y Decision Tree. Una vez generados todos los modelos y con el objetivo de generar mejores resultados en términos de precisión de la clasificación, se utiliza también un modelo de Ensemble Learning, específicamente de Stacking. Además, se genera un modelo de clasificación de texto utilizando procesamiento de lenguaje natural (Natural Languange Processing) y posteriormente un algoritmo de Machine Learning. Se verificará cual de todos estos modelos ofrece los mejores resultados en relación con la precisión de clasificación.

Naive Bayes:

El modelo de Machine Learning de Naive Bayes es un clasificador probabilístico que utiliza el teorema de Bayes que hace la suposición de independencia entre sus características. Entre sus beneficios se tiene una mayor velocidad de procesamiento al utilizar características discretas lo que a su vez genera que no sea tan eficiente con las características continuas. El modelo probabilístico utiliza la probabilidad condicional de las variables dependientes de la data usada para el entrenamiento del modelo (Awad, 2018). Se utiliza la técnica de var smoothing: 0.053 y cross validation con K fold: number splits: 5, number repetitions: 3, random state: 999.

Neural Networks:

Las redes neuronales (neural networks) en el contexto de Machine Learning son una rede de ecuaciones matemáticas que se encuentran interconectadas y reciben cierta información como input (variables) y a través del movimiento dentro la red genera variables output con el resultado esperado. Las conexiones de las redes pueden ser tratados como nodos que facilitan la comunicación entre las ecuaciones. Estos nodos se van activando a medida que son necesarios en la función general de la red. A mayor cantidad de nodos dentro de la red, la dificultad del modelo

y el tiempo de ejecución de este aumenta de manera drástica (Berwick, 2015). Se utilizaron los siguientes hiperparámetros específicos: alpha= 1, max iterations= 1000, activation= selu, epochs= 53.70, Learning rate= 0.037, neurons = 40. 87, optimizer = 0.73.

SVM (Support Vector Machine):

El modelo de Machine Learning SVM por sus siglas en inglés de Support Vector Machine realiza una división en el espacio del atributo a través de un hiperplano. Básicamente lo que hace es intentar maximizar el espacio entre las diferentes clases existentes. Se generan los vectores de soporte y se encuentra una solución óptima (Berwick, 2015). Se utilizaron los siguientes hiperparámetros específicos: Kernel: Lineal y polinomial, grado= 3, C= 1, gamma=2.

Decision Tree:

Al hablar del modelo de árbol de decisión, este se llama así puesto que genera ramas (nodos de decisión) y hojas (nodos de clasificación). Divide la data en set homogéneos con relación a que tan significativas son las variables input. Así, dependiendo de las características de cada registro, lo va clasificando y llevando por las distintas ramas hasta llegar a una calcificación final o nodo de hoja. Esta serie de decisiones conllevan a un registro a ser catalogado dentro de uno de los posibles resultados. Además, es el peso previo a la realización del modelo de Machine Learning conocido como Random Forest (Berwick, 2015). Se utilizaron los siguientes hiperparámetros específicos: max depth= 5, criterion = gini, min simple leafs= 50,

Random Forest:

El modelo Random Forest puede ser catalogado como un modelo de Ensemble Learning debido a que utiliza varios árboles de decisión para entregar un solo resultado más ajustado en términos de precisión en la clasificación. Se genera un árbol de decisión de cada muestra que se desprende del data set de entrenamiento. Se genera un data set aleatorio de atributos y como

resultado se elige el mejor atributo. Para elegir el resultado final se genera una votación del mejor árbol en relación con los demás árboles del bosque (Berwick, 2015). Se utilizaron los siguientes hiperparámetros específicos: n estimators = 20, max depth= 10, máximo características = auto, min samples leaf =4, min samples split= 10,

Ensemble Learning (Stacking):

Un modelo de Ensemble Learning busca generar un resultado más robusto al utilizar varios modelos de Machine Learning como input y generar un output con mejores métricas. Generalmente se utiliza la data de entrenamiento para entrenar y generar resultados con varios modelos. Esos resultados de los distintos modelos de Machine Learning son usados como input para clasificarlos y generar un resultado utilizando toda esta información previa (Aversana, 2018).

Fase 5: Evaluación

Una vez generados los modelos de Machine Learning previamente descritos, se compara los resultados en términos de precisión de la clasificación (Accuracy) y el F1 Score. Se determinan estas métricas ya que están relacionadas a la elección del tipo de error.

$$Accuracy = \frac{Verdaderos\ Negativos + Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos + Verdaderos\ Negativos + Falsos\ Positivos}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Figura 10: Fórmulas de precisión y F1 Score

Matriz de Confusión		
Descripción	Positivo	Negativo
Positivo	804	318
Negativo	150	598

Tabla 1: Matriz de confusión mejor modelo de Machine Learning

Así, es importante notar que para la empresa es más significativo identificar las clasificaciones que haya realizado el modelo como malas direcciones cuando en realidad estas

sean direcciones buenas (error tipo 2). También es significativo el identificar las direcciones que fueron clasificadas como buenas cuando en realidad son malas (error tipo 1). Sin embrago, el costo de no entregar un envío con la información correcta es mayor al costo operativo de intentar entregar un envío con dirección incompleta. Siguiendo este contexto, se miden los indicadores antes mencionados porque están directamente relacionados a los dos tipos de errores (Riggio, 2019).

Así se determina que el modelo que se ajusta mejor a las necesidades de clasificación de direcciones es el Random Forest puesto que presenta uno de los porcentajes de precisión más altos en conjunto con el modelo de Ensemble. A pesar de que el modelo de Ensemble tenga un punto porcentual más en relación con el F1 Score, se elige el Random Forest debido a que la capacidad computacional del modelo de Ensemble es más alta al tener que usar primero los resultados de todos los modelos aquí mencionados. Por otro lado, el Random Forest solo necesita realizar la clasificación. Al largo plazo, se busca implementar en el sistema de la empresa el algoritmo y en bases aún más grandes por lo que se toma en cuenta el tiempo de procesamiento y la capacidad necesaria para realizarlo.

Modelo	F1 Score	Accuracy
Neural Networks	74%	73%
SVM Linear	70%	69%
Naïve Bayes	68%	67%
Decision Tree	74%	71%
Random Forest	77%	75%
Ensemble	73%	75%

Tabla 2: *Resultados de los modelos de Machine Learning*

Además, se realiza un modelo de clasificación de texto utilizando Natural Language Processing (NLP). Primero, se realiza un pre-procesamiento del texto a analizar (campo dirección) a través de la extracción inicial de caracteres especiales y simplificando el texto lo máximo posible (tener todo en minúscula, sin puntuación y espacios innecesarios). Una vez realizado el

procesamiento inicial, se utilizan librerías específicas para importar lo que se conoce como Stopwords (palabras que no agregan valor al texto como artículos definidos e indefinidos, pronombres personales) en idioma español. Se extraen dichas palabras del texto y se procede con Stemming y Lemmatization (identificar cada palabra y transformarla en su lexema y morfema, eliminar cualquier tipo de prefijo o sufijo). Una vez que cada palabra este en su forma más básica, se realiza el proceso de tokenization (convertir cada oración o palabra en un token). Una vez realizado el proceso de tokenization, se realiza la vectorización de cada palabra para convertir la información de texto en información numérica que pueda ser alimentada a un modelo de Machine Learning. Se utiliza el método de vectorización de TF-IDF (Term Frequency- Inverse Document Frequencies) que básicamente genera una cuenta de los vectores en cada registro en donde el valor de una palabra se incrementa proporcionalmente a la cuenta total de todos los registros, pero a su vez es inversamente proporcional a la frecuencia de la palabra en el registro que se está analizando. Una vez realizado todo este proceso con el texto inicial, se puede entrenar al modelo de Machine Learning con data numérica. Se utiliza un modelo de Naive Bayes para realizar la clasificación (Scott, 2019). El resultado final está por debajo del mejor modelo de Machine Learning utilizando Feature Engineering por lo que se sigue eligiendo al modelo de Random Forest como el mejor.

Modelo	F1 Score	Accuracy
Random Forest	77%	75%
Text Classification	67%	65%

Tabla 3:*Resultados del mejor modelo de Machine Learning vs clasificación de texto*

Fase 6: Despliegue

Con los resultados obtenidos en el mejor modelo de Machine Learning se puede identificar que el 75% (1403) de los registros del set de prueba (20% del total de registros: 1870) fueron clasificados correctamente, sean estos positivos (direcciones correctas) o negativos (direcciones

incorrectas). Sin embargo, existe el 25% (468) de los registros que fueron clasificados erróneamente. Es importante identificar el tipo de clasificación errónea que se realiza. Así, se puede apreciar que en la matriz de confusión se tiene 318 registros de casos falsos positivos, es decir que se clasificaron como buenas direcciones cuando en realidad eran malas direcciones (Error tipo I). Por el otro lado se tiene 150 registros de casos falsos negativos, es decir direcciones clasificadas como malas cuando en realidad eran buenas direcciones (Error Tipo II). El error tipo I significaría sacar a ruta 318 registros que necesitaban de más información para poder ser entregadas efectivamente mientras que el error tipo II significa no sacar a ruta 150 registros que debían salir inmediatamente y que tenían toda la información necesaria para ser entregados. Identificar el error tipo II y minimizarlo resulta primordial para la empresa puesto que el costo de no cobrar una posible entrega efectiva ($5,00) es más grande que el costo de una parada no efectiva por información incompleta de la guía ($3,12).

Una vez que se haya validado el porcentaje error y se haya generado una mejora en la precisión de clasificación del modelo, se podría seguir con un plan de implantación conservador y progresivo en ciertos clientes o proyectos dependiendo del volumen. Se empezaría con 10% de clientes de Entregas certificadas, validar e identificar que no existan anomalías. En caso de no existir problemas, se podría seguir con un 15% adicional de clientes hasta llegar a un 25%. Una vez validado con el 25% de clientes, se debería socializar el proyecto con los involucrados del mismo, documentar el nuevo proceso y aplicarlo en más clientes. Al recibir la base de datos por parte del cliente, se debería correr de manera automática el algoritmo en el sistema, identificar los registros con direcciones correctas que pueden salir a ruta inmediatamente. Dichos registros salen a ruta en el proceso normal. En el caso de los registros con direcciones incorrectas, se debería obtener información adicional de los mismos antes de sacar a ruta y evitar generar ese costo

operativo asociado a las paradas no efectivas de los couriers. Se generó un análisis económico para identificar cual es el costo de oportunidad de la empresa al enviar a ruta registros con direcciones no clasificadas apropiadamente. Se toma el volumen de las no entregas por difícil localización del año 2020 y año 2021 y se multiplica por el 75% de precisión de clasificación del modelo (sin tomar en cuenta el 25% de error del modelo actual). Lo que se cobra por entrega se ha mantenido en el 2020 y 2021 ya que va atado a tarifas descritas en contratos. El costo por entrega va variando mes a mes dependiendo del tamaño de la operación y el volumen. Así, se saca un promedio del 2020 y lo que va del 2021. Se multiplica el volumen al 75% de cada año (restado el 10% que se podría realizar gestión interna para poder entregar) por el costo operativo por parada y se obtiene el costo por paradas no efectivas del año.

Del volumen de no entregas ya multiplicado por el 75% de precisión del modelo (tomando en cuenta el 25% de error del modelo actual) se multiplica por el 10% asumiendo de manera conservadora que de lo no entregado ese 10% podría ser gestionado internamente por la empresa a través de Call Center, mensajes de WhatsApp y consulta de bases de datos internas y externas. De esta forma se obtiene el revenue por entregas. Así, el costo de oportunidad es simplemente la suma del costo de paradas no efectivas más el revenue de entregas y el porcentaje del salario del colaborador que ya no va a tener que realizar el proceso de manera manual.

Año	No entregas	No entregas 75%	Precio por entrega	Costo Operativo por parada	Costo paradas no efectivas	Revenue Entregas	Salario colaborador	Costo Oportunidad Total
2020	50101	37575,75	$ 5,00	$ 3,12	$ 105.512,71	$ 18.787,88	$ 360,00	$ 124.660,58
2021	45658	34243,5	$ 5,00	$ 3,20	$ 98.621,28	$ 17.121,75	$ 390,00	$ 116.133,03
				TOTAL	$ 204.133,99	$ 35.909,63	$ 750,00	**$ 240.793,61**

Tabla 4:Cálculo del costo de oportunidad con la precisión actual del modelo

CONCLUSIONES Y RECOMENDACIONES

Se debe desarrollar más el modelo y disminuir el 25% de error del mismo. Una vez mejorado el porcentaje de precisión, se puede utilizar el plan de implantación descrito en la fase 6 de despliegue de la metodología CRISP-DM. El utilizar el modelo tal y como está ayudaría a generar ahorro en términos de evitar que el material que no debería salir a ruta llegue a la zona courier y revenue con lo que se logre identificar como dirección incorrecta y se gestione efectivamente información adicional. Así, el costo de oportunidad en el 2021 es de $116 133, 03 incluyendo el porcentaje del sueldo del colaborador que se dedica a esta tarea.

El utilizar Feature Engineering resultó ser un pasó fundamental en el desarrollo del proyecto, permitió obtener nueva información de la variable dirección. Esta información se transformó en nuevos campos en la base de datos que permitieron entrenar a los distintos modelos de Machine Learning. Sin esto no hubiera sido posible desarrollar los algoritmos, la base de datos inicial compartida por la empresa no contenía toda la información necesaria para el entrenamiento de los modelos.

La clasificación automática de direcciones liberará de tiempo valioso al colaborador encargado del proceso y permitirá que se cumpla al 100% en todas las bases de datos asegurando el correcto desarrollo del proceso operativo. El colaborador podrá dedicarse a otras tareas que añadan a valor en el flujo del proceso operativo.

Se recomienda que, una vez implementado el modelo, se sigan realizando pruebas y un monitoreo constante para identificar oportunidades de mejora o futuras implementaciones que añadir al modelo para hacerlo aún más completo. Se puede analizar otras áreas de la empresa que se puedan beneficiar del modelo, se pueden realizar cambios o adaptarlo a una necesidad específica de otro

departamento diferente a operaciones. También se puede empezar a recopilar estadística de cada cliente para denotar la importancia de captar bien las direcciones en un primer lugar.

Limitaciones: La base de datos inicial fue una limitante que se superó gracias al uso de Feature Engineering que permitió generar nueva información para alimentar los modelos de Machine Learning en base a la variable objetivo. Esta fue la mayor limitación puesto que en lo que respecta al acceso a dicha información no existió ningún inconveniente.

Referencias Bibliográficas:

Azevedo, A., Santos, M. (2008) KDD, SEMMA and CRISP-DM: a parallel overview. IADIS European Conference on Data Mining. Amsterdam, Países Bajos.

Awad, M. (2015) Support Vector Machines for classification. Research Gate. Recuperado de https://www.researchgate.net/publication/300723807_Support_Vector_Machines_for_Classification/link/5b5ed9b1458515c4b25271f7/download el miércoles 10 de noviembre del 2021

Banco de Desarrollo de América Latina (2016). Perfil Logístico de América Latina (PERLOG): Perfil Logísitico de Ecuador. Corporación Andina de Fomento.

Bosjank, Z. (2009) . CRISP-DM as a Framework for Discovering knowledge in small an medium sized enterprises. Department of Business Informations Systems, University of Sad. Subotica, Servia.

Comisión Económica para América Latina y el Caribe. (2019). Logística para la producción, la distribución y el comercio. Boletín 369. Facilitación, Comercio y Logística en América Latina y el Caribe. ISSN: 1564-4227

Chapman, P., Clinton, J, et. al (2000). CRISP-DM 1.0: Step by Step data mini guide. CRISP-DM Consortium. NCR Systems Engineering, Copenhagen, Noruega.

Gidudu, A. et. al. (s.f). Classification of images using Support Vector Machines. Department of Electrical and information Engineering. University of the Witwatersrand. Johannesburg, South Africa.

Goel, A. et. al. (2016). Real Time Sentiment Analysis using Naïve Bayes. Next Generation Computing Technologies. Dehadun, India.

Kullkarni, A. (2016). Random Forest Algorithm for Land Cover Classification. University of Texas at Tyler. School of Computer Science and Technology. Texas, USA.

Luo, X. (2020). Efficient text classification using selected Machine Learning techniques. Human Univesity of technology and business. China

Makkar, S. et. al. (2020). Applications of machine learning techniques in supply chain. optimization. CMR Institute of technology, Hyderabad, India.

Mazareanu, E. (2021). Size of the global logistics market in 2020, by region. Transportation and Logistics. Statista.Hamburg, Germany.

Pro-Ecuador Negocios sin Fronteras (2016). Infraestructura Logística. Ecuador. Recuperado de https://www.proecuador.gob.ec/infraestructura-logistica/ el 2 de noviembre del 2021

Riggio, C. (2019). What´s the deal with Accuracy, Precision, Recall and F1. Towards Data Science. Recuperado de https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021 el 2 de noviembre del 2021.

Sabah, M. et. al (2013). Classifying Unsolicited Bulk e-mail (UBE) using Python Machine Learning Techniques. International Journal of Hybrid Information Technology Vol. 6 No. 1. Lakehead University, Ontario, Canada.

Scott, W. (2019). TF-IDF from scratch in python: What is TF-IDF?. Towards Data Science. Recuperado de https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089 el 2 de noviembre del 2021.

Tipping, A., Kauschke, P. (2016) . PWC: Shifting Patterns, The Future of the Logistics industry. Strategy Department, USA.

Zhang, A. & Casari, A. (2018) Feature Engineering for Machine Learning. Principles and Techniques for Data Scientists. O Reilly. Boston, United States.