PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

DESARROLLO DE UN MODELO PREDICTIVO DEL TIEMPO DE ESPERA DE CAMIONES EN PUNTO DE DESCARGA DE MINERAL DE UNA UNIDAD MINA SUPERFICIAL CHILENA UTILIZANDO MACHINE LEARNING

Tesis para obtener el título profesional de Ingeniero de Minas

AUTORES:

Nicolás Lucas Rojas Telles

Sebastián Fabrizio Bravo Obregón

ASESORA:

Maribel Giovana Guzmán Córdova

Lima, Noviembre, 2024

Informe de Similitud

Yo, Maribel Giovana Guzmán Córdova, docente de la Facultad de Ciencias e Ingeniería de la Pontificia Universidad Católica del Perú, asesora de la tesis titulada “Desarrollo de un modelo predictivo del tiempo de espera de camiones en punto de descarga de mineral de una unidad mina superficial chilena utilizando machine learning”, de los autores Nicolás Lucas Rojas Telles y Sebastián Fabrizio Bravo Obregón, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 19%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 20/11/2024.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha: Lima, 19 de Noviembre de 2024.

<table>
<tr><td colspan="2">Guzmán Córdova Maribel Giovana</td></tr>
<tr><td>DNI: 08681397</td><td rowspan="2">Firma  </td></tr>
<tr><td>ORCID: 0000-0002-7954-7679</td></tr>
</table>

RESUMEN

El presente trabajo busca desarrollar un modelo predictivo que permita tener un mejor control en los procesos de carguío y acarreo dentro de una unidad minera superficial utilizando herramientas de Machine Learning.

La minería, al no poder influir en el precio de los metales (producto de venta), tiene como uno de sus objetivos principales la regulación de costos. Las actividades que generan la mayor cantidad de costos son el carguío y el acarreo; por tanto, se debe tener un monitoreo de estos procesos para así controlar y no exceder en los costos.

Actualmente, en las diferentes industrias, se viene implementando distintos modelos predictivos basados en Machine Learning mediante diversas metodologías. Este trabajo se basa en el desarrollo de un modelo predictivo aplicando la metodología desarrollada por IBM (CRISP-DM) en la industria minera. Esta consta de cinco pasos fundamentales: entendimiento del negocio, entendimiento de la data, preparación de la data, modelamiento e implementación del modelo. En el presente trabajo, solo se realizará los primeros tres pasos, puesto que realizar la implementación del modelo desarrollado no está dentro del alcance de la tesis.

Como primera etapa de la metodología, Entendimiento del Negocio, se analizó la data, proporcionada de manera confidencial, y se detectar distintas variables para una posible mejora. De todas estas variables se escogió el tiempo de espera en el punto de descarga ya que presentaba mayor variabilidad y era posible hacer una mejora a partir de un modelo predictivo porque se contaba con la data. Una vez elegida esta variable se continua con los demás pasos de la metodología hasta finalmente obtener el modelo predictivo entrenado y evaluado. En el presente trabajo se probarán tres algoritmos: Random Forest, Regresión Logística y Adabbost.

Los resultados presentados por los algoritmos mostrados fueron buenos en precisión según los indicadores obtenidos en el proceso de evaluación. Sin embargo, se deben optimizar más los modelos para obtener resultados más acertados en los otros indicadores como sensibilidad y especificidad.

ABSTRACT

Mining is a business which does not have control over sales prices since the value of metals is set by the market. For this reason, one of the primary objectives for all mining companies is to control costs. Consequently, you must have control of all the processes included in the mining cycle. The processes that represent the highest percentage of costs in mining cycle are loading and hauling. In that sense, this work aims to improve the productivity and costs of these processes through a predictive model to reflect a reduction in costs throughout the mining value chain.

Today, there are different companies from various sectors that use Machine Learning tools through a wide variety of methodologies; however, in Peruvian mining units it is not a common tool.

This research follows the CRISP-DM methodology since it is supported by a company with vast experience in the field (IBM). Through this methodology, the aim is to develop a predictive model in the hauling process of a Chilean surface mining unit.

As the first stage of the methodology, Business Understanding, the data was analyzed, and different variables were detected for possible improvement. Of all these variables, the waiting time at the discharge point was chosen since it presented greater variability and it was possible to make an improvement from a predictive model because the data was available. Once this variable has been chosen, the other steps of the methodology were performed until finally obtaining the trained and evaluated predictive model.

In the present work, three algorithms will be tested: Random Forest, Logistic Regression and Adabbost.

The development of this model will not only allow predicting the target variable: queueing time at the mineral dumping point, but also, it will be possible to visualize a general panorama of what is happening in the process to make decisions that allow achieve what was planned and not generate negative impacts on the expected costs or productivity.

Tabla de Contenido

Índice de Tablas

Índice de Gráficas

1. INTRODUCCIÓN

Una operación minera tiene como objetivo principal poder obtener el mayor beneficio económico de la extracción de materias primas de un yacimiento minero. Esto, sumado al hecho de que esta industria está sujeta a la volatibilidad de los precios de venta de los minerales, conllevan a que las empresas estén en una constante búsqueda de optimización de sus procesos que permitan el incremento de las productividades y reducción de los costos asociados. Por ello, se requiere la identificación de los principales cuellos de botella de la operación para evaluar iniciativas y oportunidades que permitan alcanzar y/o mejorar los objetivos establecidos.

Los subprocesos dentro de una operación minera que representan el mayor porcentaje de costos operativos son el carguío y acarreo con alrededor de un 50%, en promedio. Por ende, poder identificar oportunidades dentro de estos 2 subprocesos, significaría una optimización considerable dentro de los costos generales de la operación que permitirían obtener mejores indicadores económicos para la compañía.

Este trabajo se centra en desarrollar un modelo predictivo para el subproceso de acarreo para lo cual se utilizarán herramientas de Machine Learning. En la actualidad existen distintas metodologías disponibles; sin embargo, en esta investigación se optó por seguir la metodología CRISP-DM, ya que es respaldada por una empresa con vasta experiencia en el tema (IBM). Como primera etapa de la metodología, Entendimiento del Negocio, se analizó la data, proporcionada de manera confidencial, y se detectaron distintas variables para una posible mejora. De todas estas variables se escogió el tiempo de espera en el punto de descarga en chancado ya que presentaba mayor variabilidad y era posible hacer una mejora a partir de un modelo predictivo porque se contaba con la data disponible. Una vez elegida esta variable se continua con los demás pasos de la metodología hasta finalmente obtener el modelo predictivo entrenado y evaluado. En el presente trabajo se probarán tres algoritmos: Random Forest, Regresión Logística y Adaboost.

El desarrollo de este modelo no solo permitirá predecir la variable que se tenga como objetivo: tiempo de espera en el punto de descarga de mineral, sino también, se podrá visualizar un panorama general de lo que está sucediendo en el proceso para tomar decisiones que permitan lograr lo planificado y no generar impactos negativos ni en los costos ni en la productividad previstos.

1.1. Justificación

El presente trabajo se basa en el desarrollo de un modelo predictivo a ser aplicado en una operación minera mediante el uso de herramientas de Machine Learning. Estos modelos se consideran innovadores dentro de la industria minera en el Perú, ya que se encontró poca evidencia y/o antecedentes de su uso en alguna operación minera peruana.

Estudios previos internacionales han demostrado que el desarrollo de estos modelos predictivos conlleva a un mejor control de los procesos al que se les aplique. En ese sentido, al estimar lo que podría suceder en el proceso, se podría manipular las variables para llegar a cumplir con lo planificado y no incurrir en un incremento de costos.

Finalmente, esta investigación es un gran aporte para una operación minera superficial que busque tener un mejor control de sus procesos de carguío y acarreo.

1.2. Objetivos

1.2.1. Objetivo General

Desarrollar un modelo predictivo con algoritmos de Machine Learning que permita tener un mejor control de los procesos de carguío y acarreo en una operación minera a tajo abierto.

1.2.2. Objetivos Específicos

- Entender y diagnosticar el proceso de carguío y acarreo en la operación minera para la cual se desarrollará el modelo.
- Realizar un Análisis Exploratorio de Datos que permita: entender los datos, completar datos, imputar, detectar outliers.
- Manipular las variables: Ingeniería de variables y dar importancia a las variables predictoras o covariables.
- Desarrollar y evaluar el modelo: entrenar, predecir y testear el modelo creado.

1.3. Hipótesis

El desarrollo de un modelo predictivo basado en herramientas de Machine Learning ayudaría a tener un mejor control sobre los procesos de carguío y acarreo en una operación minera. Además, permitiría poder monitorear estos procesos con la finalidad de llegar a cumplir un objetivo planificado.

1.4. Alcance

Este trabajo se basa en el desarrollo del modelo predictivo con herramientas de Machine Learning, mas no en su implementación; además, se enfoca en el tipo de minería a tajo abierto y específicamente en los procesos de carguío y acarreo.

1.5. Plan de Trabajo

En esta sección del informe se describirá brevemente la metodología utilizada para alcanzar los objetivos antes mencionados.

1.5.1. Entendimiento del Negocio

Tabla 1 - Entendimiento del Negocio

Tarea	Entregable
Definición del proceso de carguío y acarreo identificando variables, actores y oportunidades de mejora	Diagrama del proceso de carguío y acarreo
Definición del objetivo o meta a alcanzar con el proyecto e identificación de métricas clave	Documento con objetivos y criterios de falla/éxito
Definición de terminología relevante en el proyecto, con miras a documentar el proyecto para futura replicabilidad	Glosario
Análisis de impacto del proyecto (¿Por qué es importante el presente proyecto?)	Documento con análisis de impacto

1.5.2. Entendimiento de la data

Tabla 2 – Entendimiento de la data

Tarea	Entregable
Descripción de los datos	Tipo y volumen de datos
Exploración de los datos (EDA)	Gráficos e Informes
Análisis de calidad de datos: identificar outliers, errores, vacíos y proponer posibles soluciones	Informe de calidad de datos

1.5.3. Preparación de la data

Tabla 3 – Preparación de la data

Tarea	Entregable
Selección de variables relevantes. Criterios de selección apuntando al objetivo (*target*)	Listado de variables con su respectivo puntaje
Limpieza y Transformación de datos: rectificación de valores perdidos, estandarización, manejo de *outliers, encoding* de variables categóricas	Bases de datos lista para realizar el modelo predictivo

1.5.4. Desarrollo del Modelo

Tabla 4 – Desarrollo del Modelo

Tarea	Entregable
Selección de modelos a probar argumentando en base a criterios de negocio y al objetivo	Documento argumentativo de los modelos escogidos
Describir la metodología de testeo del modelo y partición de los datos en entrenamiento y prueba para validar el modelo	Metodología de testeo y partición de data
Construcción del modelo, ajuste de parámetros	Jupyter Notebook con el desarrollo del modelo
Evaluación del modelo en base a un testeo de datos: validación del modelo	Matriz de confusión y precisión del modelo

1.5.5. Evaluación en términos mineros

Tabla 5- Evaluación del modelo

Tarea	Entregable
Evaluación de los resultados obtenidos con respecto al objetivo de negocios. Impacto de resultados en el negocio	Documento con los resultados obtenidos

1.6. Antecedentes

1.6.1. Deep Neural Network for Predicting Ore Production by Truck-Haulage Systems in Open-Pit Mines

Se propone utilizar una red neuronal profunda, una herramienta de *Machine Learning*, para crear un sistema predictivo que ayude a estimar la producción diaria de una mina a cielo abierto. Esta necesidad surge debido a que las operaciones mineras son sistemas muy dinámicos y sus entornos de trabajo son muy cambiantes día a día. Por tal motivo, se desarrolló un sistema predictivo efectivo, el cual fue entrenado con una data lo suficientemente amplia en periodo de tiempo, para que esta contase con la mayor cantidad de situaciones posibles. La data utilizada para la creación de este modelo fue tomada de un periodo de 2 meses, con la cual se realizaron 2 modelos predictivos, uno para el turno de día y otro para el turno de noche. Finalmente, con los resultados obtenidos a partir de estos 2 modelos, se tuvo un error promedio de cerca del 4.17%. Adicionalmente, el autor menciona que es importante que el sistema tenga una actualización constante con respecto a la data para que el modelo pueda seguir aumentando y/o manteniendo su efectividad, esto debido a los ya mencionados entornos de trabajo cambiantes. (Baek and Choi 2020)

1.6.2. Simulation of Truck Haulage Operations in an Undeground Mine Using Big Data from an ICT-Based Mine Safety Management System

Se busca presentar los resultados de un trabajo de investigación en el cual la data proveniente del sistema de seguridad de una mina subterránea es utilizada para la simulación de los sistemas de carguío y acarreo en esta operación minera. El método utilizado fue un análisis estadístico de cerca de 600 operaciones de carguío y acarreo. Dichos resultados fueron comparados con procesos actuales de carguío y acarreo para 2 días del mes de enero de 2019. Por un lado, en el caso de las toneladas totales producidas, la diferencia fue de 30 y 0 toneladas para cada uno de los días. Por otro lado, con respecto a los tiempos de viaje, la diferencia promedio fue de 0.13 y 0.14 minutos. Con esto se concluyó que el sistema obtenido es capaz de poder predecir la producción de los camiones, la producción total de la operación y los tiempos para cada una de las etapas en este proceso. (Baek and Choi 2019)

1.6.3. Modelo predictivo basado en machine learning de órdenes de trabajo riesgosas para mantenimiento de equipos mineros

Se desarrolló un modelo predictivo que permita tener una gestión proactiva sobre las

órdenes de trabajo de mantenimiento. Esto se debe a que, en la empresa donde fue realizado este estudio, cerca del 15% de las órdenes de trabajo no eran desarrolladas adecuadamente. Por tal motivo, lo que se buscaba era introducir toda la data existente dentro del modelo predictivo para que, de esta manera, se puedan analizar e identificar patrones históricos de fallas a nivel de materiales y cómo estas se encuentran relacionadas con otras variables que también tienen cierto grado de influencia dentro del manejo de las órdenes de los materiales. En tal sentido, a la finalización de este proyecto de investigación, se pudo obtener el modelo predictivo el cual estaba basado en el uso del algoritmo de Gradient Boosting Trees. Este, luego de ser analizado con operaciones actuales, tuvo una precisión de cerca del 82%, con lo cual se concluye que el sistema creado es efectivo en el proceso de predicción y automatización del análisis de las órdenes para operaciones futuras. (Barroso Salgado 2019).

2. MARCO TEÓRICO Y METODOLOGÍA

2.1. Carguío y Acarreo en Minería Superficial

El principal objetivo de cualquier operación minera es extraer minerales de la corteza terrestre de una manera factible en términos económicos, ambientales y sociales. Este proceso requiere de un análisis exhaustivo que permita identificar cuáles serían los métodos óptimos a emplear para poder obtener el mayor retorno económico posible. En consecuencia, cada uno de los factores involucrados en este proceso tendrán una importancia considerable para la elección del método de minado a emplear y cuáles serían las condiciones en las cuales cada una de las etapas del ciclo de minado operarán.

2.1.1. Ciclo de minado en Minería Superficial

El ciclo de extracción minera para el tipo de operación a la cual se encuentra dirigido nuestro proyecto, minería superficial, consta de 4 principales procesos: perforación, voladura, carguío y acarreo. Además, la existencia de ciertas operaciones auxiliares que podrían estar presentes durante el proceso dependerá de las características y condiciones de operación del proyecto.

2.1.2. Proceso de carguío y acarreo

La importancia de los procesos de carguío y acarreo radica en el hecho de que ambos procesos, en conjunto, constituyen alrededor del 50% de los costos operativos, en promedio, de una operación minera. Por tal motivo, un adecuado análisis de estos y de los factores que influyen en su efectividad, permitirán reducir los costos operacionales totales.

2.1.2.1. Carguío

El carguío es una de las operaciones unitarias pertenecientes al ciclo de minado de toda operación minera. Este consiste en el proceso de carga de material fragmentado una vez finalizado el proceso de voladura del material. Se realiza mediante el uso de equipos de carguío, usualmente palas electrohidráulicas o cargadores frontales, hacia camiones para el posterior proceso de acarreo.

Para poder lograr que la operación de carguío se realice de una manera eficiente, se debe considerar los principales factores que afectan directa e indirectamente. Algunos de estos factores son los siguientes:

- Densidad del material
- Factor de esponjamiento del material

- Características de los equipos a utilizar
- Dimensiones de las labores
- Condiciones del lugar de carguío

Con los factores mencionados, lo que se busca es realizar un análisis efectivo de estos en conjunto que permita lograr el máximo provecho en cuanto a la utilización de los equipos y el rendimiento que estos puedan brindar. Para ello, se realizará un cálculo de la cantidad de equipos necesarios a utilizar y el tiempo estimado por cada ciclo de carga.

2.1.2.2. Acarreo

Una vez finalizado el proceso de carguío de material dentro de los camiones, se procede al proceso de acarreo de dicho material hacia su respectivo destino (PADs de Lixiviación, Chancadora, Pilas de mineral o botaderos de estéril Este es el proceso de mayor complejidad dentro de las operaciones ya que resulta un reto poder realizar una selección óptima de la combinación de camiones para poder lograr una adecuada optimización de los costos y tiempos asociados de este proceso.

El principal objetivo de este proceso es definir cuáles son los sectores de carga, las direcciones de carguío (a frentes de carga, posición de equipos de carguío y nivel de pisos) y el destino de los materiales de acuerdo con leyes de clasificación y tonelajes definidas previamente. (Barreto Taipe 2017)

Los constantes cambios en los puntos de carguío, así como la existencia de distintos puntos de destino dependiendo de las características del material, son algunas de las condiciones que pueden variar dentro la operación minera. En tal sentido, un sistema de análisis que nos permita estimar y reducir cada uno de los ciclos dentro del proceso de acarreo resultará efectivo en términos económicos para la operación. Dichos ciclos se encuentran clasificados de la siguiente manera: (Infante 2019)

- **Tiempo viajando vacío:** Tiempo en la cual el camión se encuentra dirigiéndose hacia la excavadora (punto de carguío) desde el lugar en el que acaba de descargar material.
- **Tiempo Esperando:** Debido a retrasos operativos, el camión tiene tiempos en los cuales se encuentra esperando su turno para poder ser cargado en el área de carguío.
- **Tiempo Cuadrando:** Inicia cuando el camión llega al área de carga y procede a colocarse de una manera adecuada para proceder a ser cargado.

- **Tiempo Cargando:** Después de haberse cuadrado, inicia el tiempo en el cual el camión está siendo cargado por los equipos de carguío.
- **Tiempo Acarreando:** Tiempo en el cual el camión se encuentra acarreando el material desde el área de carga hasta su destino.
- **Tiempo en Cola:** Debido a retrasos operacionales, el camión se encuentra con otros camiones dentro del área de descarga, lo cual ocasionará que se tenga un tiempo de espera.
- **Tiempo Retrocediendo:** Inicia cuando el camión se coloca en una posición adecuada para efectuar la descarga del material.
- **Tiempo Descargando:** Tiempo que inicia cuando el camión comienza la acción de descargar material en el área de descarga.

Lo que se busca es reducir al máximo cada uno de estos tiempos individualmente para lograr tener una operación con mayor efectividad.

2.1.3. Indicadores de Gestión de Flota en los procesos de Carguío y Acarreo

Una forma efectiva de poder medir el grado de cumplimiento de los objetivos establecidos en el ciclo minero es mediante el empleo de los KPI´s o Indicadores Claves de Rendimiento. Estas pueden ser mediciones financieras o no financieras que reflejan su rendimiento en un periodo determinado y son empleados en diversas áreas de una empresa.

Dentro de una operación minera estos son utilizados para determinar el estado actual del negocio y permiten tomar acciones correctivas en función de la línea de acción futura. Esto con el objetivo de poder incrementar el desempeño favorable para conseguir el progreso, posicionamiento y ganancias significativas de la empresa. Algunos de los principales KPI's utilizados en la industria son los KPI´s de productividad. Entre los más utilizados se tiene:

- **Disponibilidad Mecánica (DM):** Por disponibilidad mecánica se entiende el porcentaje del tiempo total en el que el equipo se encuentra disponible para operar. El porcentaje en el cual el equipo no se encuentra disponible se debe principalmente a paras por mantenimientos programados y no programados. Por tal motivo, este indicador de rendimiento es manejado por el área de mantenimiento. (José Eder Bustamante Chávez 2018)

$$Disponibilidad\ Mecánica = \frac{Horas\ Totales - Horas\ Malogrado}{Horas\ Totales}$$

- **Uso de la disponibilidad (UD):** Uso de la disponibilidad hace referencia al tiempo en el cual el equipo se encuentra encendido, en producción o en demoras. Este tiempo es medido respecto al tiempo en el que se encuentra mecánicamente disponible; es decir, al punto mencionado anteriormente.(José Eder Bustamante Chávez 2018)

$$Uso\ de\ la\ disponibilidad = \frac{Horas\ Operativas + Demoras}{Horas\ Totales - Horas\ Malogrado}$$

$$Uso\ de\ la\ disponibilidad = \frac{Horas\ Operativas + Demoras}{Horas\ Operativas + Demoras + Horas\ Stand\ By}$$

- **Uso del equipo (Use):** De igual forma, el uso del equipo es el tiempo en el cual un equipo se encuentra produciendo respecto al tiempo en cual el equipo se encuentra con el motor encendido. Una forma práctica de poder entender este parámetro es mediante las demoras a las cuales están sometidas los distintos equipos. En un capítulo posterior se detalla en profundidad cuáles son los principales tipos de demoras y su causa.(José Eder Bustamante Chávez 2018)

$$Uso = \frac{Horas\ Operativas}{Horas\ Operativas + Demoras}$$

- **Utilización:** Este se mide en porcentaje y nos refleja el tiempo en el cual el equipo se encontraba operativo respecto al tiempo total disponible para el equipo. (José Eder Bustamante Chávez 2018)

$$Utilización = \frac{Horas\ Operativas}{Horas\ Totales}$$

- **Tiempo Medio Entre Reparaciones (MTTR):** MTTR (Mean Time To Repair, por sus siglas en inglés), nos indica el tiempo promedio por reparación. Este nos indica la mantenibilidad de los equipos y la eficacia de dichos mantenimientos. (José Eder Bustamante Chávez 2018)

$$MTTR = \frac{Horas\ en\ Reparación}{Número\ de\ paradas}$$

- **Tiempo Medio Entre Fallas (MTBF):** Este parámetro indica el tiempo promedio entre paradas por mantenimiento y nos indica la confiabilidad y eficiencia del mantenimiento. (José Eder Bustamante Chávez 2018)

$$MTBS = \frac{Horas\ Trabajadas}{Número\ de\ paradas}$$

- **Productividad:** Indicador de rendimiento que establece las toneladas producidas en un intervalo de tiempo determinado. (José Eder Bustamante Chávez 2018)

$$Productividad = \frac{Toneladas\ Producidas}{Intervalo\ de\ tiempo}$$

- **TKPHr**: Establece las toneladas producidas en una distancia promedio y en un intervalo de tiempo.

$$TKPHr = \frac{Toneladas\ producidas * Distancia\ promedio}{Intervalo\ de\ tiempo}$$

- **Dig rate:** Indicador utilizado por equipos de carguío y establece la velocidad de minado de estos.

$$Dig\ rate = \frac{Toneladas\ producidas}{Intervalo\ de\ tiempo}$$

- **Queue time:** Este nos indica el tiempo en el que el equipo se encuentra esperando su turno para poder continuar con el ciclo. (José Eder Bustamante Chávez 2018)

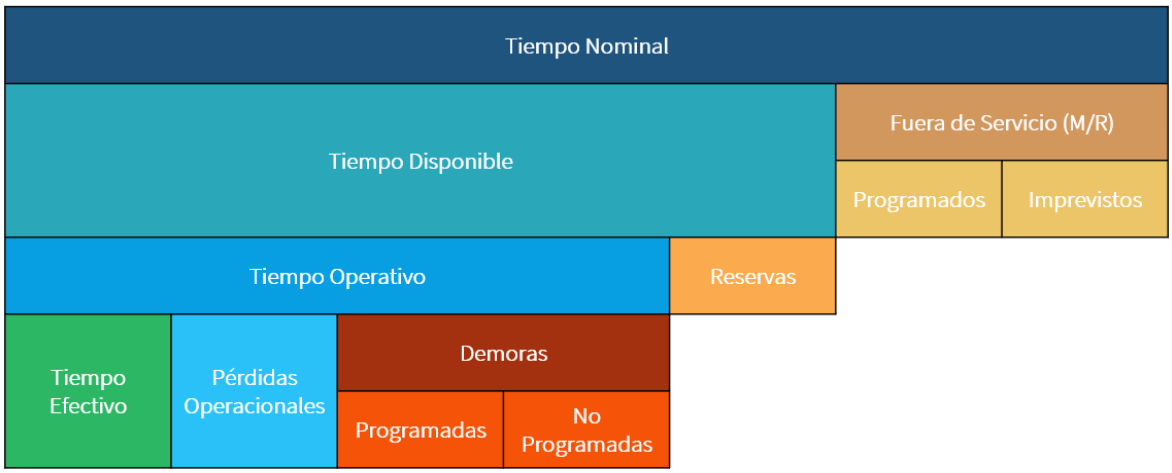
$$Queue\ time = Tiempo\ total\ de\ espera * \frac{(x\ horas)}{(x\ minutos)}$$

2.2. Conceptos Mineros

2.2.1. Sistema ASARCO de gestión de flota

En el capítulo 2.1.3 de los KPI's se mencionó la importancia de poder calcular los diversos Índices Operacionales para la operación minera. En tal sentido, una forma habitual en la que se puede tener un control y evaluación efectiva de gestión de los estatus operacionales es mediante el uso de la escala de tiempos según la norma ASARCO. Esta norma clasifica y describe en detalle cada uno de los estatus en los que los equipos se encuentran en un periodo determinado de tiempo durante la operación. (Bonzi 2016)

Tabla 6: Distribución de tiempos del modelo ASARCO



- **Tiempo Nominal:** Esto hace referencia a todo el tiempo en el cual el equipo se encuentra físicamente en faena.
- **Tiempo Disponible:** Tiempo en el cual el equipo se encuentra en óptimas condiciones y habilitado para poder operar. Del Tiempo Nominal se descuenta el tiempo en el que el equipo se encuentra en las mantenciones.
- **Tiempo no Disponible:** Este hace referencia al tiempo en el cual el equipo se encuentra recibiendo algún tipo de mantención y/o reparaciones. Estas pueden estar clasificadas como Mantenimiento Programado y Mantenimiento No Programado.
- **Mantenimiento Programado**: En este se considera todo el tiempo en el cual el equipo se encuentra recibiendo algún tipo de mantenimiento, el cual ha sido previamente programado por el área correspondiente. Esta puede estar clasificada como Mantenimiento Preventivo (Se realiza cuando se observa que algún componente del equipo se encuentra próximo a fallar) y Mantenimiento Planificado (Mantenimiento que se encuentra supeditado por el cronograma inicial de mantenimiento que es generado antes del inicio de las operaciones para cada uno de los equipos).
- **Mantenimiento No Programado:** Se realiza como medida de corrección ante un imprevisto generado por la descomposición de algún equipo.

Es importante mencionar que cada empresa minera dispone de un propio sistema de clasificación en el que se pueden incluir nuevos o quitar otros parámetros.

2.2.2. Sistema DISPATCH de gestión de flota

Como dentro de cualquier proceso minero, es necesario el empleo de un sistema que pueda realizar seguimiento y control en tiempo real de todos los equipos y parámetros dentro

de este. El objetivo principal es que estos se controlen para mantenerlos dentro de los valores establecidos previamente para el cumplimiento de los estándares operacionales. Por tal motivo, en la actualidad, la mayoría de las operaciones mineras que buscan alcanzar dichos objetivos cuentan con un sistema de despacho. (José Eder Bustamante Chávez 2018)

El sistema de despacho Dispatch es un software diseñado para ser utilizado por empresas mineras como herramienta de gestión de flota. Este cuenta con una gran aceptación dentro del sector, puesto que ha demostrado ser de gran utilidad debido a los resultados ampliamente satisfactorios obtenidos durante más de 20 años de uso. (Vidal 2010)

Este sistema se basa en un algoritmo que busca optimizar la asignación de camiones a palas y/o botaderos y/o chancado (destino), maximizando la utilización del tiempo y minimizando las pérdidas mediante el registro de información de estos procesos en tiempo real. El sistema Dispatch® tiene como objetivos principales:(Bonzi 2016)

- Automatizar y optimizar asignaciones de camiones
- Archivar datos para equipos de carguío, transporte y auxiliares
- Asignación de combustible automáticamente
- Recolector de datos para mantenimiento
- Mezclar minerales
- Reportabilidad propia y a través de Power View® según la necesidad del cliente
- Aumentar la productividad
- Reducir costos de operación (Bonzi 2016)

Su sistema de operación se basa principalmente en una herramienta que registra en tiempo real los eventos que se desarrollan durante los ciclos de operación de los procesos de carguío y acarreo. Con dicha información, el algoritmo se encarga de generar de forma automática la ruta óptima de acarreo. Las operaciones básicas de este sistema son: (Bonzi 2016)

1. Registro de eventos del ciclo de acarreo de material relevantes e importantes para la empresa
2. Traspaso de los datos en tiempo real y posterior decodificación
3. El sistema registra los datos y guarda la información
4. El software procesa los datos y gestiona asignaciones de manera óptima de destino a los camiones de extracción
5. Finalmente, el sistema hace un envío de la asignación al camión respectivo para que realice la ruta óptima.

Modelo minero:

El sistema Dispatach cuenta con un modelo minero cuyas variables deben de ser configuradas en las minas donde este sea implementado. Dichas variables y su configuración, mediante el uso del algoritmo presente, optimizan la gran cantidad de información y variables que entran al sistema. De esta forma, lo que se busca es que el sistema provea la mejor asignación a cada uno de los equipos en cuestión en el menor tiempo posible. La configuración que debe de ser introducida acorde a cada mina en específico incluye las siguientes variables: (Bonzi 2016)

- Mina: Ubicación geográfica en el espacio en el que está siendo explotada
- Pit: Lugar físico de explotación donde se encuentran los equipos operando
- Región: Un pit siempre se encontrará separado por sectores con la finalidad de poder puntualizar la operación en sectores específicos.
- Punto de Carga: Ubicación donde se realiza la operación de carguío, usualmente es la ubicación de los equipos de carguío
- Punto de Descarga: Ubicación donde se realiza la operación de descarga por parte de los camiones, usualmente es la ubicación correspondiente a botaderos, stocks y/o chancadoras.
- Nodos intermedios: Dentro del sistema son puntos virtuales definidos por coordenadas especiales. Estos son utilizados para referenciar caminos; es decir, entre cada nodo se puede conocer la distancia y pendiente de los caminos.
- Balizas: Son puntos virtuales dentro de las rutas de acarreo e incluyen puntos de carga y descarga dentro del pit y se definen por coordenadas espaciales. Estos sirven para poder detectar los camiones de acarreo en sus llegadas y/o salidas desde el origen o destino. En estos puntos, al sistema se le permite reasignar la ruta de los camiones si es que se da el caso de que encuentre una mejor.

2.2.3. Diagrama de Pareto

Como bien se sabe, dentro de cualquier proceso existen un sinnúmero de variables que tienen un impacto dentro de estos; sin embargo, no todos impactan en la misma proporción. Una de las principales formas de poder identificar dichas variables es mediante el uso del Principio de Pareto. Este consiste en la identificación de la importancia de ciertas variables y el énfasis de las de mayor importancia, variables vitales, sobre las variables de menor importancia, variables triviales. Propone, además, que de las variables identificadas, no todas pueden ser controladas y necesitan ser identificadas como tal y ser separadas de las que sí se puede aplicar algún método correctivo. (José Eder Bustamante Chávez 2018)

Algunas de las principales ventajas que se obtienen al utilizar esta herramienta se mencionan a continuación:

- Indica la causa principal de los problemas y permite establecer la prioridad en la cual deben de ser resueltos.
- Representa en forma ordenada la ocurrencia del mayor al menor impacto de los problemas o áreas de oportunidad de mejora
- Permite identificar oportunidades de mejora
- Analizar distintas agrupaciones de datos
- Para evaluar los resultados de los cambios realizados a un proceso (evaluar la evolución)

2.3. Ciencia de Datos

La ciencia de datos es la intersección de tres aspectos: Ciencias computacionales, estadística y dominio en un tema específico. De las ciencias computacionales provienen el Machine Learning y las tecnologías informáticas de alto rendimiento. De la estadística, el tradicional análisis exploratorio de datos (EDA), pruebas de significancia y visualización de datos. Por último, del dominio del tema, las problemáticas a analizar y la evaluación de los resultados producto de la ciencia de datos (Skiena 2017).

La novedad de la ciencia de datos no se basa en los últimos conocimientos científicos, sino en un cambio disruptivo en nuestra sociedad que ha sido causado por la evolución de la tecnología: “Datificación”. La datificación es el proceso de convertir en datos aspectos del mundo que nunca se han cuantificado antes. (Igual and Seguí 2017)

Existen tres principales motivos por los que la Ciencia de Datos se ha vuelto un tema muy utilizado últimamente(Skiena 2017):

- Las nuevas tecnologías permiten capturar, anotar y almacenar grandes cantidades de datos desde distintas fuentes como la nube. Después de haber acumulado una buena cantidad de datos, empiezas a preguntarte qué aporte puedes generar con ellos.
- Los avances en las ciencias informáticas permiten analizar datos de formas novedosas y en escalas cada vez mayores. Las arquitecturas de computación en la nube brindan incluso acceso a un gran poder cuando lo necesitan.
- Grandes empresas destacadas han demostrado que la Ciencia de Datos genera una variedad de beneficios de la analítica de datos.

2.3.1. Modelos Predictivos

Los modelos predictivos son métodos que permiten detectar patrones y tendencias en los datos recopilados de distintas fuentes. Utilizando estos modelos, las empresas pueden detectar problemas o detectar nuevas oportunidades. Es por ello, que hoy en día, el análisis predictivo se ha vuelto muy común en las distintas empresas alrededor del mundo ya que este les brinda una gran herramienta al momento de tomar decisiones. Esto se debe a que, en base a patrones repetitivos encontrados en la data histórica, se puede actuar con un sustento más teórico. Por otro lado, un modelo predictivo se podría también automatizar para analizar la data en tiempo real(Barroso Salgado 2019). Sin embargo, el modelo que se realizará en este trabajo solo será desarrollado, mas no implementado para su automatización. Un ejemplo de un modelo predictivo es el sistema de recomendación de las distintas empresas como Netflix o Facebook que básicamente utilizan distintas fuentes de data para inferir los intereses del cliente. (Aggarwal 2016)

2.3.2. Machine Learning

El Machine Learning es un tema muy conocido en los últimos años. Este es una forma de inteligencia artificial que permite a una máquina o sistema aprender de un set de datos en lugar de aprender mediante programación explícita. (Kubat 2017)

A pesar de ser muy conocido, el Machine Learning, no es un proceso sencillo. Un modelo de Machine Learning es el producto del entrenamiento de un algoritmo con la información dentro de la data. Luego del entrenamiento, al darle un nuevo set de datos al modelo, este generará un pronóstico basado en los datos que entrenaron al modelo. Además, es muy importante tener en cuenta que conforme el algoritmo ingiere mayor cantidad de datos de entrenamiento, es posible producir modelos más precisos basados en datos (IBM 2020a). Por lo tanto, es muy importante garantizar la calidad y la cantidad de los datos de entrada para crear un modelo que tenga la capacidad predictiva deseada.

Dentro del Machine Learning existen varios enfoques de aprendizaje que dependen del problema a abordar, de los tipos y de los volúmenes de los datos:

2.3.2.1. Aprendizaje Supervisado

Este tipo de aprendizaje tiene como punto de partida un conjunto establecido de datos y la variable objetivo, definida dentro del set de datos. Según la naturaleza de la variable objetivo, este tipo de aprendizaje se puede subdividir en: Regresión, si la variable objetivo es cuantitativa o, Clasificación, si es categórica. Por ejemplo, se puede crear un modelo que

determine si una persona caerá o no en morosidad, si se entrenase al modelo con un set de datos que contenga un historial de personas morosas o no morosas en función a distintas variables como la edad, género, etc.

2.3.2.2. Aprendizaje No supervisado

El aprendizaje no supervisado se utiliza cuando el problema requiere una cantidad masiva de datos sin etiquetar; es decir, no se tiene dentro del set de datos la variable objetivo. La comprensión del significado detrás de estos datos requiere algoritmos que clasifican los datos con base en los patrones o clústeres que encuentra.

El aprendizaje no supervisado lleva a cabo un proceso iterativo, analizando los datos sin intervención humana. Un ejemplo podría ser la tecnología de detección de spam en e-mails donde los clasificadores de Machine Learning, basados en Clustering y asociación, se aplican para identificar e-mail no deseado. (IBM 2020a)

2.3.3. Python

Python es en leguaje de programación que lidera en el desarrollo de Machine Learning. Los principales motivos son su simplicidad y facilidad de aprendizaje. Es utilizado por grandes desarrolladores de modelos hasta por principiantes en Machine Learning.(Gonzales 2020)

Python no tiene integrados el procesamiento de datos o las expresiones matemáticas científicas dentro del lenguaje en sí, sin embargo, existen librerías como Scipy, Numpy y Pandas que ofrecen dichas funcionalidades de una manera más accesible. Además, existen librerías especializadas en Machine Learning como Scikit-learn, Theano, TensorFlow. Estas le dan al desarrollador la capacidad de crear una gran variedad de modelos utilizando Python.(Gonzales 2020)

Como lenguaje de programación ha ganado interés en los últimos años, particularmente en el mundo comercial, ya que ha aumentado el número de personas que desea aprender Python. Este crecimiento de interés se debe a distintos factores como la flexibilidad y facilidad de aprendizaje de este, la cantidad de librerías (módulos) existentes y que es de libre uso (sin licencia alguna). (Hunt 2019)

2.3.4. Librerías de Python utilizadas

2.3.4.1. Librerías para la Ciencia de Datos

Para el manejo de datos en Python se utiliza distintas librerías que facilitan el análisis y la comprensión de los datos. Las librerías mencionadas a continuación, son una herramienta de gran utilidad ya que las otras librerías especializadas se basan en estas.

- **SciPy**

Esta es una librería utilizada tanto en ciencia de datos como en ingeniería. SciPy incluye dentro de sus módulos específicos distintas funciones para optimización, integración, estadística y algebra lineal; se basa en Numpy y al agregar una gama de algoritmos y comandos de alto nivel se logra manipular y visualizar los datos.(Gonzales 2020)

- **Numpy**

Numpy significa Numerical Python. Esta librería es fundamental dentro de Python, ya que proporciona la vectorización de las operaciones matemáticas en matrices. Así se logra mejorar el rendimiento y acelerar la ejecución. Su propósito principal es proporcionar la capacidad de hacer operaciones complejas de matrices. Esto hace que operaciones requeridas por redes neuronales o estadística compleja se puedan realizar de manera más rápida y sencilla. Numpy es una librería que está emparejada a otras librerías orientadas a Machine Learning como Pandas, Matplotlib, entre otras. (Gonzales 2020)

- **Pandas**

Es una librería de Python especialmente creada para trabajar con datos “etiquetados” y “relacionales” de manera simple e intuitiva. Pandas facilita la manipulación, agregación y visualización de datos. El elemento principal es el DataFrame. Esta librería proporciona herramientas para configurar, fusionar, remodelar y dividir grandes bases de datos; además, otra ventaja es que se puede convertir estructuras de datos en objetos y manejar datos faltantes (NaN). (Gonzales 2020)

2.3.4.2. Librerías para la Visualización de Datos

Además de hacer el análisis de los datos, es muy importante reflejarlo mediante distintas gráficas. Esto ayudará a poder entender mejor los datos y a transmitir el análisis a otras personas. Para esto, dentro de Python existen librerías especializadas en crear fácilmente gráficos, tablas y mapas atractivos y sofisticados.

- **Matplotlib**

Es una librería estándar de Python que permite crear diagramas y gráficos. Dentro Matplotlib existen módulos que ayudan a generar gráficos más avanzados. La ventaja principal de Matplotlib es la flexibilidad y la gran cantidad de comandos para hacer cualquier tipo de gráfico que se desee. (Gonzales 2020)

2.3.4.3. Librerías para Machine Learning

Existen librerías que incluyen dentro de sus módulos distintas funcionalidades que facilitan el desarrollo de modelos predictivos basados en Machine Learning. Por ejemplo, se puede hacer la separación de base de datos, el entrenamiento y la evaluación del modelo. Además, estas librerías ya incluyen los algoritmos que se utilizarán para crear el modelo. La librería más utilizada y conocida dentro de la ciencia de datos con Machine Learning es Scikit-Learn.

- **Scikit-learn**

Esta librería es probablemente una de las más populares de Machine Learning. Tiene una gran cantidad de utilidades para la minería de datos y el análisis de datos. Scikit-learn está construido sobre las bases de otras librerías populares ya mencionadas como Numpy, Scipy y Matplotlib. (Gonzales 2020)

Scikit-learn incluye dentro de sus módulos una gran cantidad de algoritmos de Machine Learning tanto supervisado como no supervisado. Además, permite evaluar mediante distintas metodologías los modelos creados como por ejemplo mediante validación cruzada.

Una característica muy importante de esta librería es que existe gran cantidad y calidad de documentación. Esto permitirá realizar consultas de cualquier problema que se presente.

2.3.5. Algoritmos de Machine Learning

Cuando se comienza el desarrollo de un modelo predictivo no es posible saber de antemano qué algoritmo es el adecuado para el problema. Por tanto, se deben probar distintos métodos, evaluarlos y centrarse en aquellos que tengan mejor rendimiento y resultados. En el presente trabajo se utilizará aprendizaje supervisado para clasificar y determinar si se cumplirá o no el tonelaje a fin de turno. En consecuencia, en esta sección se describirán algoritmos de Machine Learning utilizados para la clasificación que serán probados y, finalmente, se escogerá el mejor durante el desarrollo de la presente tesis.

2.3.5.1. Regresión Logística

Este algoritmo es un método predice clases binarias. La variable objetivo puede ser de dos clases posibles. Por ejemplo, se puede usar para determinar si se cumple el tonelaje a fin de turno o no. Este algoritmo es uno de los más usados para la clasificación de dos categorías. Esta estima la relación entre las variables independientes con la variable binaria dependiente. (Gonzales 2020)

2.3.5.2. Adaboost

AdaBoost es un algoritmo de machine learning de boosting que se aplica comúnmente a la clasificación. Fue creado por Yoav Freund y Robert Schapire en 1996. Básicamente, la idea detrás de AdaBoost es aumentar clasificadores débiles para obtener unO fuerte. A continuación, se muestra una descripción general del algoritmo.

1. Inicialización: Comienza con un conjunto de datos de entrenamiento y asigna un peso igual a cada observación.
2. Entrenamiento Iterativo:
    - **Primer Clasificador Débil**: Se entrena un clasificador débil (por ejemplo, un árbol de decisión de una profundidad) en el conjunto de datos.
    - **Actualización de Pesos**: Las observaciones mal clasificadas por el primer clasificador reciben un mayor peso, mientras que las correctamente clasificadas reciben un peso menor.
    - **Siguientes Clasificadores Débiles**: Se entrena un nuevo clasificador débil en el conjunto de datos con los pesos actualizados. Este proceso se repite iterativamente.
3. Combinación de Clasificadores: Al final, los clasificadores débiles se combinan mediante una ponderación basada en su precisión. Los clasificadores más precisos tienen un mayor peso en la decisión final. (Aljarah, 2020)

2.3.5.3. Bosques Aleatorios para clasificación (Random Forest Classifier)

También conocidos como Random Forest, es una técnica de clasificación versátil de Machine Learning. Mediante este método, se lleva a cabo métodos de reducción dimensional, trata valores perdidos, valores atípicos y otros pasos esenciales de exploración de datos.

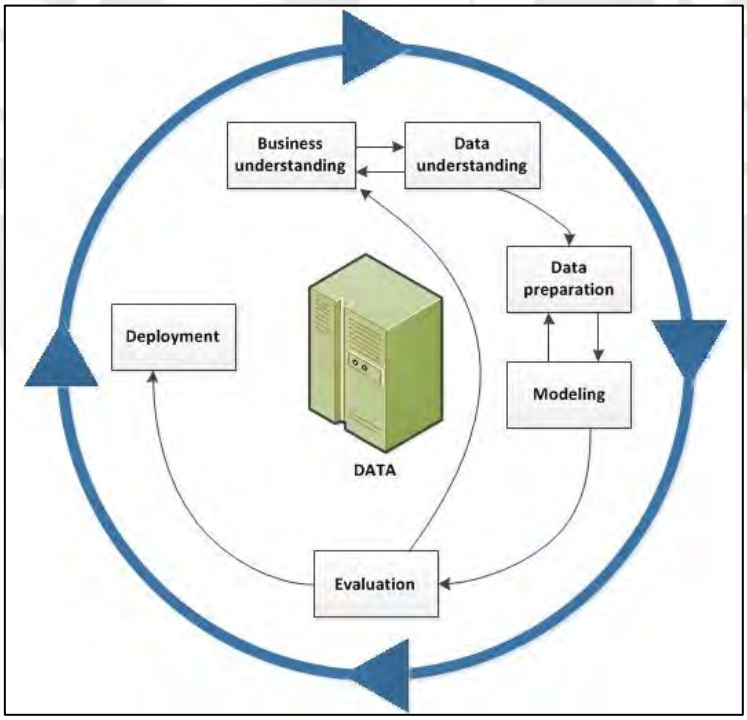
Los bosques aleatorios son una combinación de árboles predictores en donde cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente y

con la misma distribución para todos los árboles del bosque. (Pavlov 2019)

Este es un tipo de método de aprendizaje por conjuntos; es decir, un grupo de modelos débiles se combinan para formar un modelo poderoso. En los boques aleatorios se cultivan varios árboles en lugar de un solo árbol. En otras palabras, para clasificar un nuevo objeto basado en atributos, cada árbol da una clasificación y se dice que el árbol “vota” por esa clase. Finalmente, el Random Forest elige la clasificación con más votos.(Gonzales 2020)

2.4. Metodología de Desarrollo del Modelo (CRISP-DM)

El método CISP-DM (Cross Industry Standard Process for Data Mining) es un método probado desarrollado por IBM que es muy utilizado para orientar trabajos de minería de datos. Esta metodología incluye descripciones de las fases de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas. Además, como proceso, ofrece un resumen del ciclo vital de minería de datos. En el presente gráfico se esquematiza el proceso para el desarrollo de la metodología CRISP-DM. (IBM 2020b).



Gráfica 1: Esquema de la metodología CRISP-DM (Fuente: IBM)

El ciclo vital de la metodología contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases. La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario. Por otro lado, es muy importante mencionar que en el presente trabajo solo se llegará hasta la quinta fase ya que no se realizará el despliegue o implementación del

modelo. A continuación, se realizará una descripción de las fases a utilizar en el presente trabajo de tesis:

2.4.1. Entendimiento del Negocio

Durante esta fase se hace una evaluación de la situación dentro del negocio en el que se realizará el modelo. Se deben determinar: El problema, Los objetivos y Recursos disponibles (Material y Personal). Es muy importante poder expresar la eficacia del modelo en términos del negocio. (IBM 2020b) Dentro de las tareas principales en esta fase se tienen:

- Determinar la estructura de la organización
- Describir el problema
- Describir la solución actual
- Definir objetivos en términos de negocio
- Establecer criterios para evaluar el rendimiento del modelo en términos del negocio.

2.4.2. Entendimiento de Datos

La fase de entendimiento de datos de CRISP-DM implica estudiar más de cerca los datos disponibles. Este paso es esencial para evitar problemas inesperados durante la siguiente fase (preparación de datos), la cual suele ser la fase más larga de un proyecto. La comprensión de datos implica acceder a los datos y explorarlos con la ayuda de tablas y gráficos que se pueden organizar. (IBM 2020b) En este trabajo se realizará el Análisis Exploratorio de Datos (EDA) utilizando las librerías de Python ya mencionadas previamente como Pandas, Matplotlib, entre otras. Los datos se analizan para producir buenos resultados y tomar buenas decisiones sobre los datos.

EDA visualiza distribuciones de datos como gráficos de barras, histogramas, diagramas de caja. Además, permite visualizar correlaciones (relaciones) entre variables a través de mapas de calor. (Sahoo et al. 2019) Entre las tareas principales mencionamos:

- Recopilación de datos iniciales
- Descripción de los datos
- Exploración de los datos
- Verificación de la calidad de los datos

2.4.3. Preparación de Datos

Esta fase es uno de los aspectos más importantes y con frecuencia que más tiempo exige en la minería de datos. En la mayoría de los proyectos, la preparación de datos suele

llevar el 50-70 % del tiempo y esfuerzo. Dedicar los esfuerzos adecuados a las primeras fases de comprensión comercial y comprensión de datos puede reducir al mínimo los gastos indirectos relacionados, pero aún deberá dedicar una buena cantidad de esfuerzo para preparar y empaquetar los datos.(IBM 2020b) Dependiendo del trabajo a realizar y sus objetivos, la preparación de datos implica las tareas siguientes:

- Fusión de registros de datos
- Selección de una muestra de un subconjunto de datos
- Agregación de registros
- Clasificación de los datos para el modelamiento
- Eliminación o Imputación de valores nulos o perdidos: La imputación debe considerarse parte del proceso de investigación con el propósito de arribar a conclusiones sustentadas en evidencia empírica sólida.(Medina and Galván 2007)
- Limpieza de datos
- Construcción de nuevos datos: Ingeniería de Variables
- Codificación de variables categóricas
- División del set de datos en set de prueba y set de entrenamiento

2.4.4. Modelamiento

En esta etapa, es donde se concreta el trabajo realizado en las etapas previas. Los datos preparados se incorporan a las herramientas analíticas de Machine Learning que se usarán. Los resultados se aproximarán a la solución del problema planteado en Comprensión del negocio. En esta etapa, se utilizará la librería mencionada anteriormente Scikit-learn.

En el modelamiento, se suele ejecutar múltiples iteraciones. Es decir, se ejecutan varios modelos utilizando los parámetros predeterminados, luego se ajustan los parámetros o, si es necesario, se retorna a la fase de preparación de datos. No es común que los problemas planteados en la ciencia de datos presenten una única solución. Existen muchas formas para resolver un problema concreto. (IBM 2020b) Entre las principales tareas se tiene:

- Selección de algoritmos para el modelamiento
- Selección de los métodos de evaluación para los diferentes modelos
- Desarrollo o Generación del modelo
- Evaluación del modelo

2.4.5. Evaluación del Modelo

En esta fase del proyecto, se habrá completado la mayor parte del mismo. También

se habrá determinado, en la fase de modelado, que los modelos creados, en base a distintas herramientas de Machine Learning, son técnicamente correctos y efectivos en función de los criterios de rendimiento computacionales que se han definido previamente. Sin embargo, antes de continuar, se debe evaluar los resultados utilizando los criterios de rendimiento del negocio establecidos en el inicio del proyecto. Esta es la clave para asegurar que el modelo pueda utilizar los resultados que ha obtenido. (IBM 2020b) Dentro de las tareas se tiene:

- Evaluación de Resultados
- Proceso de Revisión
- Determinación de los siguientes pasos

3. DESARROLLO DEL MODELO

3.1. Descripción de la Data

3.1.1. Obtención de la data

Las tablas conteniendo la data provienen del sistema de despacho que se maneja en esta unidad minera. Las tablas proporcionadas fueron:

(1) Hist_loads: Esta tabla registra cada ciclo de carga realizado por una pala y un camión.

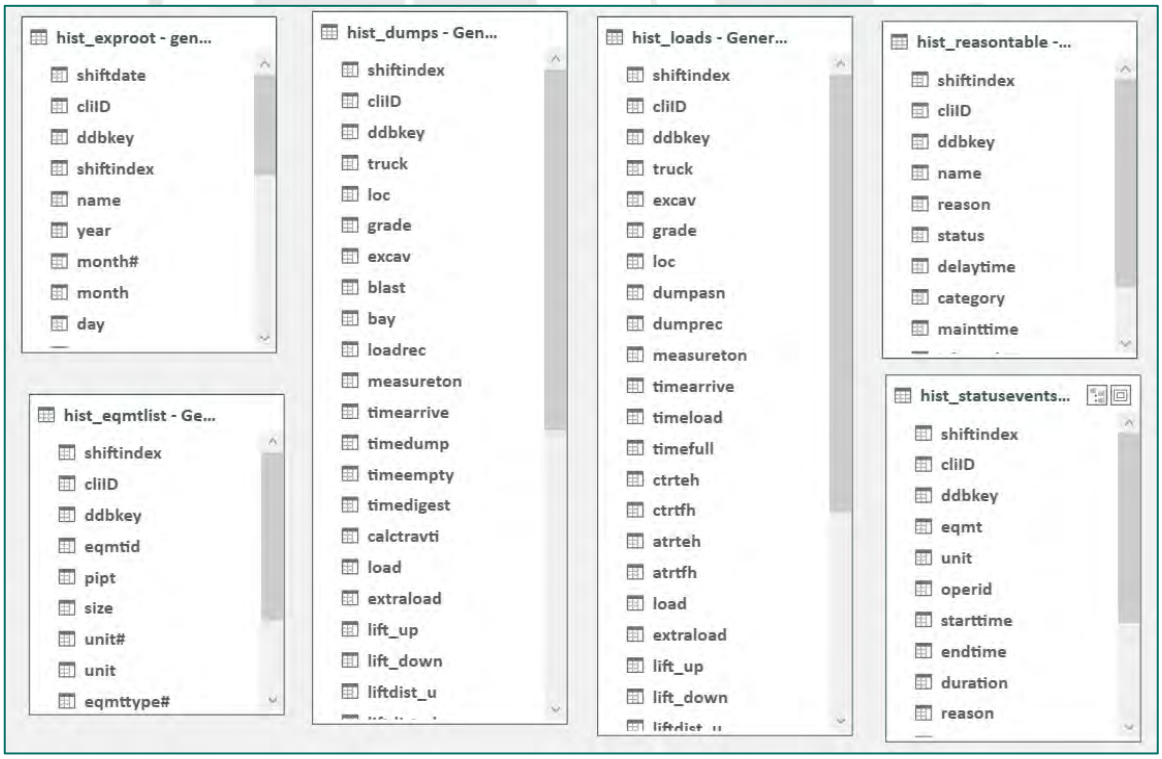
(2) His_dumps: Esta tabla contiene las descargas realizadas por los camiones y está relacionada directamente con la tabla hist_loads

(3) Hist_eqmplist: Contiene las especificaciones técnicas de cada equipo utilizado en la unidad minera

(4) Hist_exproot: Es la tabla calendario del sistema de despacho

(5) Hist_reasontable: Contiene la razón de cada estado presentado en la tabla status

(6) Hist_statusevents: Contiene la información (como duración) sobre cada estado en el que se encuentra los equipos.



Gráfica 2: Tablas proporcionadas para el trabajo

3.2. Preparación de la tabla a modelar

De las tablas proporcionadas se tuvo que seleccionar las variables que servirían para

el desarrollo del modelo. Para esto se realizaron distintas tareas como: combinación, filtrado, selección de variables, entre otras. Finalmente, la tabla a ser utilizada es la siguiente:

Tabla 7: Variables utilizadas para el modelo predictivo

Variable	Descripción de la variable
shiftindex	Índice de guardia con el que está asociado la carga
seq	Ventana de tiempo de 2 horas en la guardia (1 a 6)
cola_descarga_prom_ch	Tiempo promedio de cola en zona de chancado
dumping_time_otros_ratio	Ratio entre tiempo de descarga y tiempo operative en puntos distintos al de la zona de chancado
dumping_time_ch_ratio	Ratio entre tiempo de descarga y tiempo operative en puntos de chancado
tons_descarga	Total de toneladas alimentadas
tons_ch	Toneladas alimentadas a la chancadora
cola_pala_cargadores_ratio	Porcentaje del tiempo de cola en los cargadores sobre el tiempo de operación
cola_pala_shovel_ratio	Porcentaje del tiempo de cola en las palas sobre el tiempo de operación
spottime_cargadores_ratio	Porcentaje del tiempo de cuadrado en los cargadores sobre el tiempo operativo
spottime_palas_ratio	Porcentaje del tiempo de cuadrado en las palas sobre el tiempo operativo
cola_descarga_otros_ratio	Porcentaje del tiempo de cola en las zonas de descarga sobre el tiempo operativo
cola_descarga_ch_ratio	Porcentaje del tiempo de cola en la zona de chancado sobre el tiempo operativo
viaje_lleno_ratio	Porcentaje del tiempo de acarreo cargado sobre el tiempo operativo
viaje_vacio_ratio	Porcentaje del tiempo de acarreo vacío sobre el tiempo operativo
REND_CAEX	Productividad de los camiones
UT_CAEX	Utilización de los camiones
DF_CAEX	Disponibilidad de los camiones
cant_caex	Cantidad de camiones operativos
tons_carga	Toneladas cargadas
hangtime_cargadores_ratio	Porcentaje del tiempo de cola sobre el tiempo operative de los cargadores
loadingtime_cargadores_ratio	Porcentaje del tiempo de carga sobre el tiempo operative de las palas
hangtime_palas_ratio	Porcentaje del tiempo de cola sobre el tiempo operative de las palas
loadingtime_palas_ratio	Porcentaje del tiempo de carguío sobre el tiempo operativo de las palas
UT_CF	Utilización de los cargadores
DF_CF	Disponibilidad de los cargadores
REND_carguio	Productividad de los equipos de carguío
UT_palas	Utilización de las palas
DF_palas	Disponibilidad de las palas
cant_cargadores	Cantidad de cargadores operativos
cant_palas	Cantidad de palas operativas
target	Variable que identifica si el tiempo promedio de cola en la zona de chancado es más alto (1) o menor (O) que 86 segundos

3.3. Desarrollo de la metodología

3.3.1. Entendimiento del Negocio

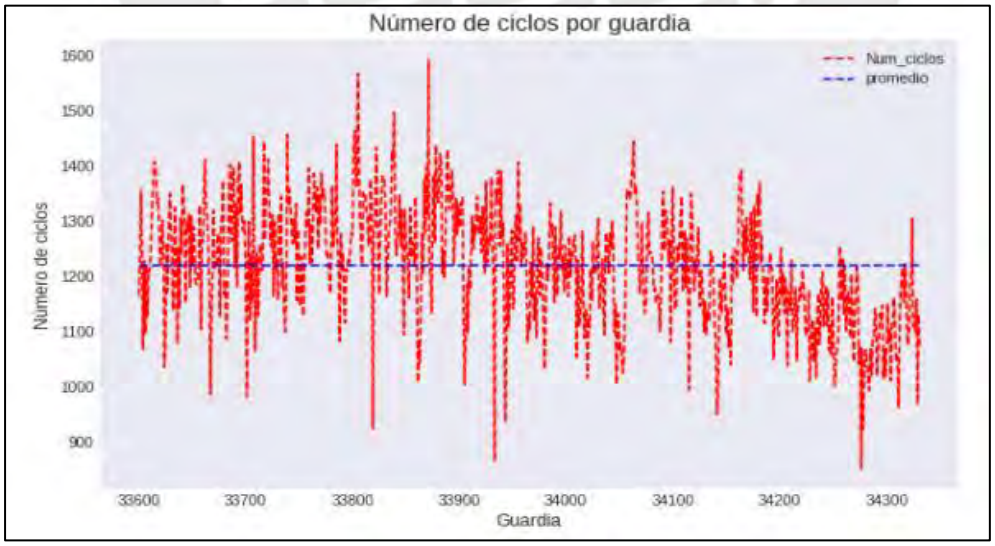
3.3.1.1. Descripción general del proyecto desde los datos

La data es proveniente del sistema de despacho de una operación minera ubicada al norte de Chile. Además, es importante recalcar la confidencialidad de los datos por lo que no se podrá revelar información específica como el nombre exacto de los puntos de descarga o de los tajos. La información general del proyecto que se pudo obtener desde las diferentes tablas se presenta en la Tabla 9.

Tabla 8: Descripción general del proyecto desde los datos

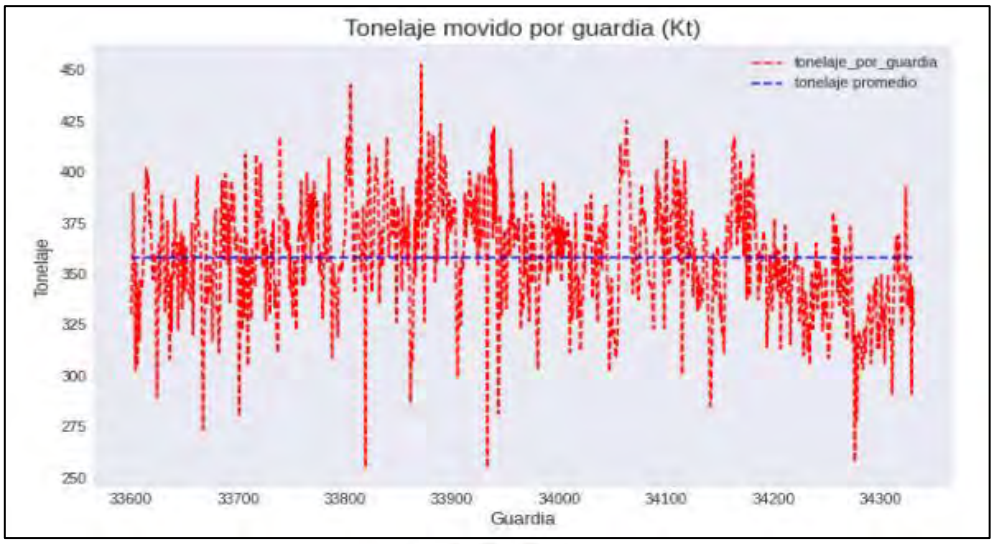
Descripción General	
Año	2016
Numero de guardias registradas	732
Guardias al día	2
Duración de la guardia (horas)	12
Numero promedio de ciclos por guardia	1220
Producción promedio por guardia	358.03
Producción en el año 2016 (Kilo toneladas)	262,047

Además, se han obtenido gráficas de línea del número de ciclos y de la producción por guardia para dar una idea del rango en el que se viene trabajando en esta operación minera.



Gráfica 3: Gráfica de líneas del Número de ciclos por guardia

En esta primera gráfica se puede observar el comportamiento del número de ciclos que se realizó en cada guardia a lo largo del año. El promedio de 1220 ciclos por guardia es apreciado en la línea azul.



Gráfica 4:: Gráfica de líneas del Número de tonelaje movido por guardia

Mediante esta gráfica se observa el comportamiento del tonelaje movido en las diferentes guardias a lo largo del año. De color azul, se puede apreciar el valor promedio en este caso 358 Kt por guardia (aproximadamente 700 Kt por día).

Otro análisis importante que se realizó es determinar la cantidad de equipos destinados para los diferentes procesos dentro del ciclo minero de esta operación. Para esto se obtuvo un diagrama de barras.

Gráfica 5: Gráfico de barras de Número de equipos usados en cada proceso del ciclo minero

En el gráfico se puede observar la gran cantidad de equipos (camiones) utilizados en el proceso de acarreo. En ese sentido, se puede demostrar la importancia de una mejora en esta etapa en el ciclo minero. Existen 113 camiones mineros y 14 equipos de carguío los cuales son los que dirigen la unidad minera en estudio.

3.3.1.2. Definición y descripción del proceso de Carguío y Acarreo en la unidad minera

a) Ciclo de Carguío y Acarreo pala-camión:

Para comprender y analizar los procesos de Carguío y Acarreo, es necesario entender los siguientes términos dentro de dichas actividades.

Los tiempos de actividad se calculan en función de los datos del Global Positioning System (GPS) determinando y el tiempo que el camión pasa en un área determinada. Las actividades dentro de este ciclo vienen definidas por: (Chaowasakoo et al. 2017)

(1) Viaje vacío: Tiempo que el camión vacío viaja hacia una pala ubicada en el área del punto de carguío.

(2) Tiempo en espera:

   a. Para camiones se refiere al tiempo que espera el camión en una pala. En esta operación lo definen como *queuetime*.
   b. Por otro lado, para palas se refiere al tiempo desde el momento en que sale el camión anterior del área de carga hasta el momento en que el próximo camión tome posición de carga. En esta mina viene definido como *hangtime*

(3) Aculatamiento: Tiempo que el camión demora para posicionarse para el cargado. En la empresa lo definen como *spottime*.

(4) Tiempo de Carga: Tiempo de carga del camión con la pala. Se refiere la diferencia entre el momento en el que se carga el primer material en el camión y el momento en que el camión sale de la zona de carga.

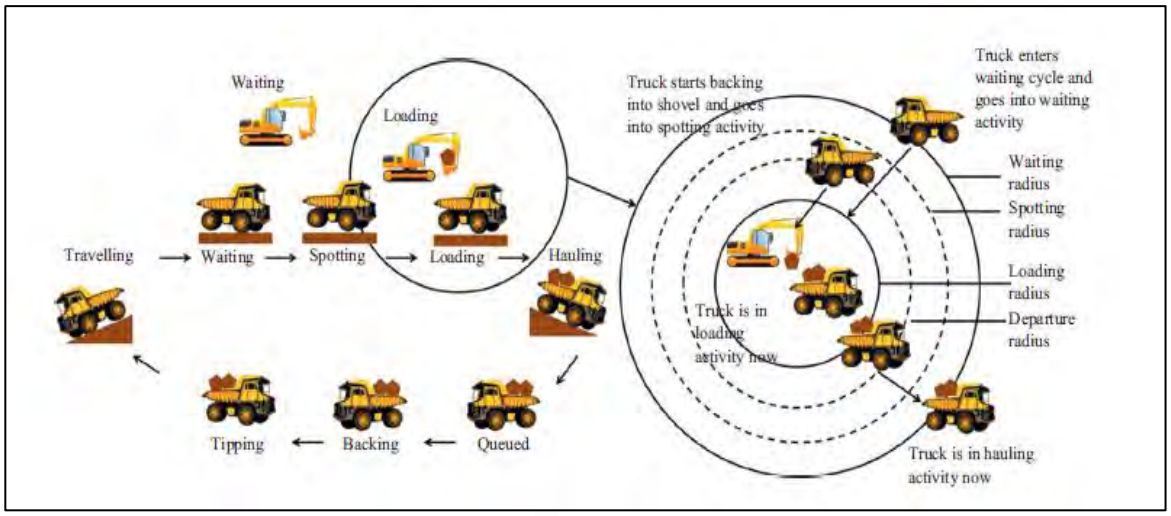
(5) Viaje Lleno: Tiempo que el camión transporta material hasta un punto de descarga.

(6) Tiempo en cola: Tiempo que el camión espera en el punto de descarga. En esta mina viene dado por *idletime.*

(7) Retroceso: se refiere al tiempo que el camión dedica a tomar la posición para la descarga.

(8) Descarga: Tiempo que el camión se toma para descargar el material

La siguiente gráfica ilustra una operación cíclica de la actividad de un camión y una pala.



Gráfica 6: Ciclo de Carguío y Acarreo pala-camión (Chaowasakoo et al. 2017)

b) Equipos utilizados en los procesos de Carguío y Acarreo

1. Proceso de Carguío

En el proceso de carguío se cuenta con un total de 14 equipos. En la siguiente tabla se presentan los cinco tipos de equipos con sus respectivas capacidades.

Tabla 9: Cantidad de Equipos de Carguío por capacidad y tipo

Flota de Carguío		
Tipo	**Capacidad (yd^3)**	**Cantidad**
Bucyrus 495 BII	56	2
Bucyrus 495 HR	73	4
KMS-PC8000	42	1
Le Tourneau L1850	36	3
P&H 4100XPC	74	4

2. Proceso de Acarreo

Para el proceso de acarreo se cuenta con un total de 113 equipos. En la siguiente tabla se muestran la capacidad y tipo de cada equipo respectivo:

Tabla 10: Cantidad de Equipos de Acarreo por capacidad y tipo

Flota de Acarreo		
Tipo	**Capacidad (ton)**	**Cantidad**
Haulpack-830E	150	2
Haulpack-830E	217	17
Haulpack-830E	240	2
Haulpack-830E	303	1
KMS-930 E3	150	3
KMS-930 E3	264	1
KMS-930 E3	300	82
KMS-930 E3	303	1
LBH T282C	373	4

3.3.1.3. Identificación de variables, actores y oportunidades de mejora en los procesos de Carguío y Acarreo

3.3.1.3.1. Análisis de Demoras

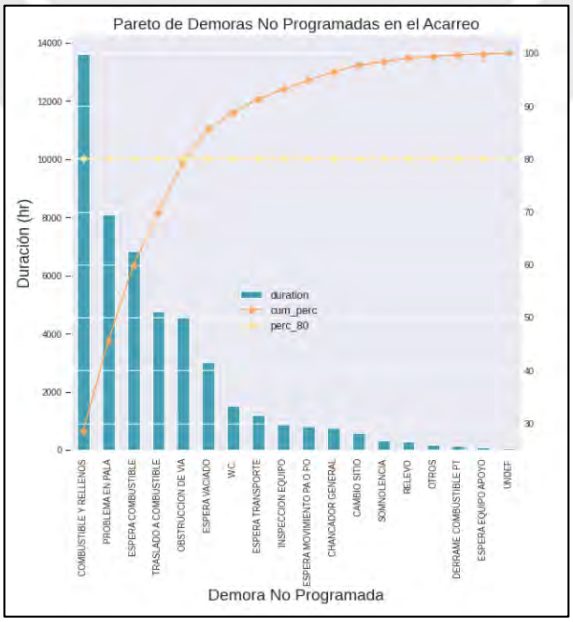
Se realizó un análisis de las demoras programadas y no programadas para los ciclos de carguío y acarreo con la finalidad de encontrar las causas más recurrentes de las demoras y las que más horas tomaban.

El diagrama de Pareto nos permitió detectar las demoras que más horas presentaban en los procesos. Así se pudo detectar oportunidades de mejora. Se dividió el análisis para cada proceso: carguío y acarreo.

Para cada proceso se presentará dos diagramas de Pareto. Uno para las demoras programadas y otro para las no programadas. Debido a la naturaleza de las demoras, se decide seleccionar variables que estén dentro del Pareto de demoras No Programadas ya que este tipo de demoras podrán mejorarse mediante una optimización operacional utilizando data adquirida. Por otro lado, las demoras programadas solo serían optimizadas con una mejora en la administración o en la logística de esta con otra fuente de datos.

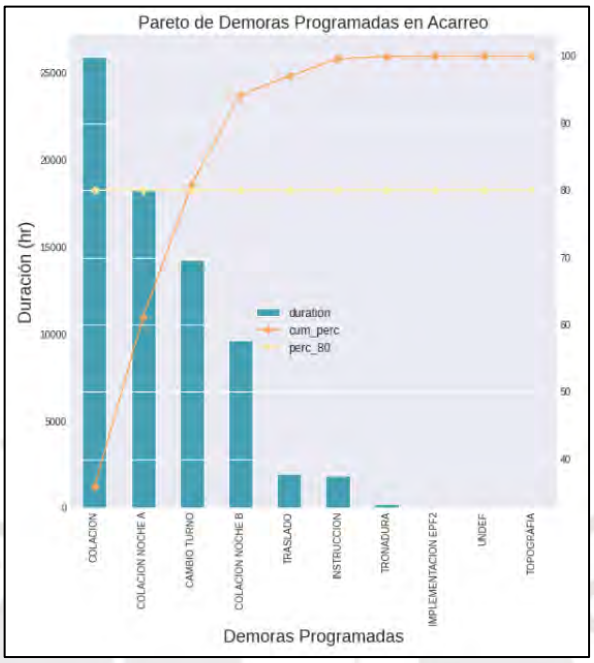
a) Acarreo:

Demoras No Programadas: En la siguiente gráfica se puede observar que las demoras que toman más tiempo son las relacionadas al combustible y la espera en el punto de descarga (espera en descarga). De este análisis solo se tomó en cuenta la espera en vaciado ya que este tipo de demora está representada en la data por una variable específica llamada *idletime.*



Gráfica 7: Diagrama de Pareto de Demoras No Programadas en Acarreo

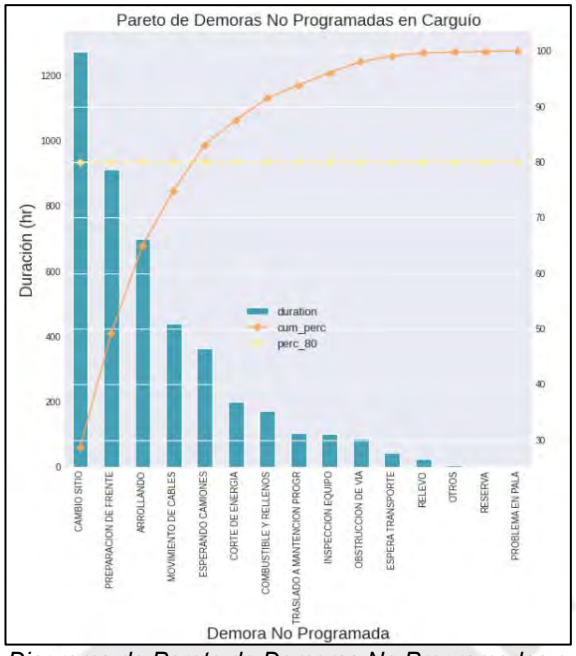
Demoras Programadas: En el siguiente diagrama se puede apreciar que las demoras programadas que más tiempo toman son las relacionadas al cambio de guardia y a la hora de refrigerio. Como se mencionó, el alcance de este trabajo no contempla ese tipo de mejoras por lo que no se escogerá ninguna de estas variables como oportunidades de mejora.



Gráfica 8: Diagrama de Pareto de Demoras Programadas en Acarreo

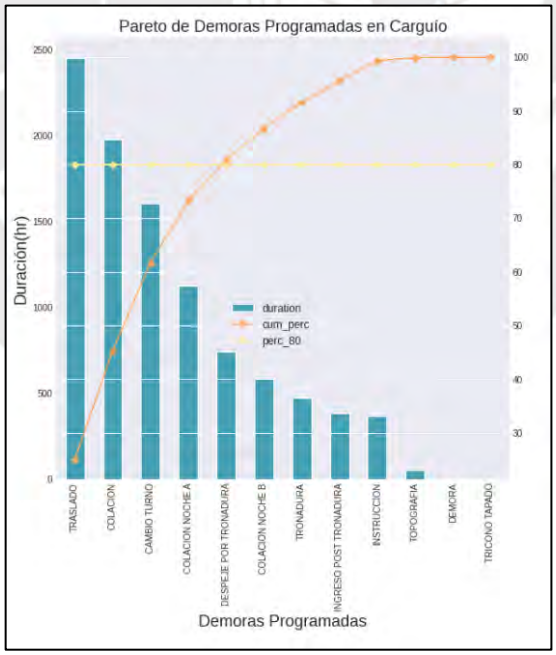
b) Carguío:

Demoras No Programadas: En la siguiente gráfica se puede observar que las demoras que más tiempo toman son las relacionadas con los servicios auxiliares como: el cambio de sitio o el tiempo arrollando. Además, una demora importante, por parte de la pala, es la espera a que llegue un camión al punto de carga. De este análisis solo se tomó en consideración la espera de la pala ya que este tipo de demora está representada en la data por una variable específica llamada hang*time.*



Gráfica 9: Diagrama de Pareto de Demoras No Programadas en Carguío

Demoras Programadas: Se puede apreciar en la Grafica 10 que las demoras programadas que más tiempo toman son las relacionadas al cambio de guardia y a la hora de refrigerio. Como se mencionó, este trabajo de tesis no está enfocado a ese tipo de mejoras por lo que no se escogerá ninguna de estas variables como oportunidades de mejora.



Gráfica 10: Diagrama de Pareto de Demoras Programadas en Carguío

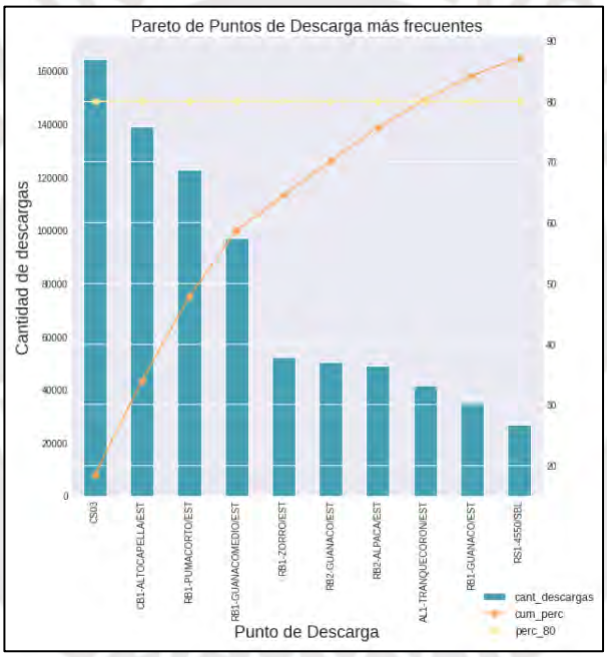
Con los análisis realizados se decidió optar por la “demora del tiempo de espera” en el punto de descarga en el proceso de Acarreo. Esto debido a que representaba una mayor cantidad de horas lo cual se traduciría en mayor tiempo recuperable al utilizar un modelo de

predicción de la variable que la representa: *idletime.* En consecuencia, esta variable seleccionada será analizada en profundidad continuando con el entendimiento del negocio para determinar en qué punto de descarga es que está presentándose mayor variabilidad, lo que representaría una falta de control y una oportunidad de mejora.

3.3.1.3.2. Análisis de Tiempo de Espera en Puntos de Descarga (Idletime)

a) Puntos de Descarga más recurrentes:

En primer lugar, mediante el Diagrama de Pareto, se determinará los puntos de descarga más recurrentes. Así se podrá analizar los puntos en los que se centra la actividad minera en la empresa analizada. Una vez obtenidos los puntos más recurrentes, se procederá a analizar cada uno de ellos por separado para determinar cuál presenta una mayor oportunidad de mejora.

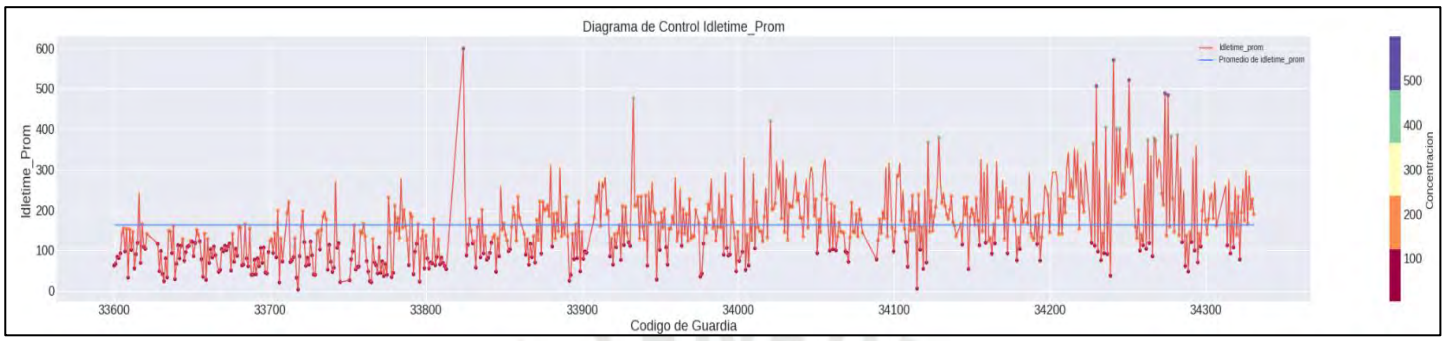


Gráfica 11: Pareto de Puntos de Descarga más frecuentes

En el gráfico 11 se puede apreciar que el punto que contempla la mayor cantidad de descargas es el punto de descarga de mineral denominado “CS03”, seguido por otros puntos de descarga de material estéril. Se analizará el punto de descarga de mineral y uno de los puntos de descarga de estéril para determinar en cual de estos se presenta una mayor variabilidad. Se espera que el punto de descarga de desmonte tenga un tiempo de espera cercano a cero ya que este cuenta con un mayor frente para realizar la actividad de descarga, por lo que no existen demoras para realizar dicha actividad. Por otro lado, probablemente el tiempo de espera sea mayor en el CS03 debido a que el espacio designado para dicha actividad es reducido. Todo esto se demostrará a través de los datos en análisis posteriores.

b) Puntos de Descarga de mineral CS03

Para poder analizar la variabilidad en este punto de descarga se examinaron diferentes ventanas de tiempo en las cuales se observó el comportamiento de la variable que representa la espera en descarga llamada *idletime*.



Gráfica 12: Diagrama de líneas de la variable idletime por guardia en punto CS03

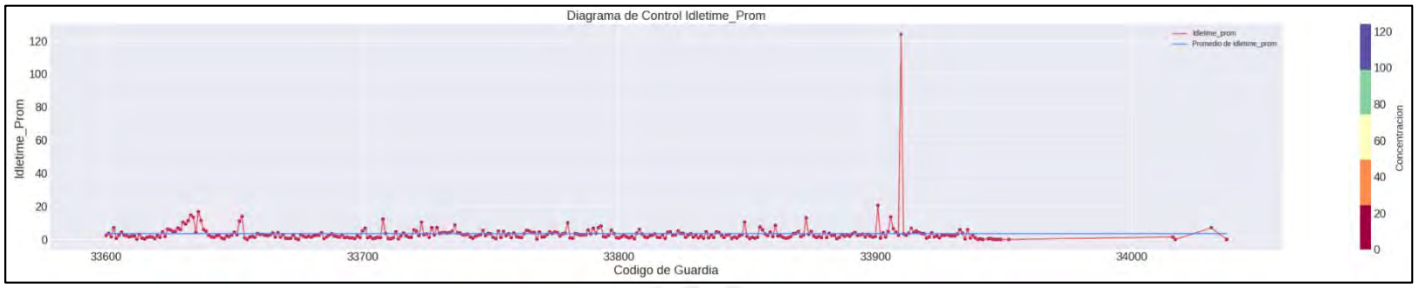
En la gráfica 12 se puede observar que la variable presenta mucha dispersión con respecto de su promedio que viene representado por la línea azul. Esto quiere decir que en este punto la variable no cuenta con un control adecuado debido a la alta variabilidad encontrada. Además, se puede ver el alto valor de la media mayor a tres minutos (aproximadamente 200 segundos).

Para profundizar más el análisis en este punto, se analizó la variable determinando su promedio por cada mes (Gráfica 13). Así, se podrá comparar posteriormente el comportamiento de esta variable en los otros puntos de descarga. En este punto de descarga se puede observar que el *idletime* ha tenido una tendencia a subir a lo largo de los meses. En el mes de noviembre se llegó al valor más alto y en diciembre se tuvo una reducción. En general, se puede decir que en este punto la variable se encuentra en valores elevados y que no cuenta con un control adecuado.

Gráfica 13: Idletime promedio mensual en punto CS03 2016

c) Puntos de Descarga de desmonte ALTOCAPELLA/EST

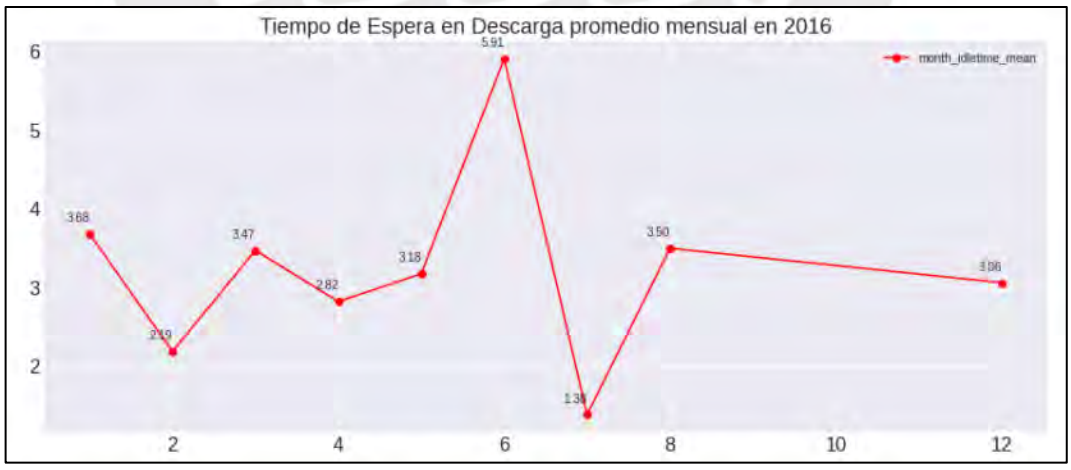
Para poder analizar la variabilidad en este punto de descarga se examinaron diferentes ventanas de tiempo en las cuales se observó el comportamiento de la variable que representa la espera en descarga llamada *idletime*.



Gráfica 14: Diagrama de líneas de la variable idletime por guardia en punto ALTOCAPELLA/EST

En la gráfica 14 se aprecia que la variable *idletime* no presenta mucha dispersión con respecto de su promedio que viene representado por la línea azul. Esto quiere decir que en este punto la variable cuenta con un control adecuado debido a la baja variabilidad encontrada. Además, se puede ver el alto valor de la media de más de tres minutos (aproximadamente 200 segundos). Por otro lado, existen puntos con valores elevados, a los cuales se les puede considerar como valores atípicos debido a una mala toma de datos u otras razones.

En el análisis del promedio mensual de la variable *idletime* (Gráfica 15) se puede ver que la variable se encuentra en valores muy bajos, lo cual nos indica que en este punto de descarga no hay motivo de demoras.



Gráfica 15: Idletime promedio mensual en punto ALTOCAPELLA/EST

Este análisis se realizó en todos los puntos de descarga de desmonte y se encontró el mismo comportamiento: valores bajos y resultados no dispersos. En ese sentido, como se esperaba y se mencionó anteriormente, se determinó que la oportunidad de mejora se encuentra en la variable idletime en el punto de descarga mineral CS03.

El siguiente paso es estimar de qué manera se puede mejorar la variable y cómo impactará la mejora en el proceso. Es decir, determinar cuántas horas se podrían recuperar y calcular, con dichas horas, cuantas toneladas adicionales de mineral se podrían producir.

3.3.1.3.3. Estimación de la pérdida en horas y en tonelaje

Como se mencionó previamente, se estimará un rango de horas optimizables, es decir, la cantidad de horas que se podrían recuperar mediante un modelo predictivo. Para esto, se definió estadísticamente un valor representado por el P90 de la variable. De este modo, se determinó que los valores de *idletime* mayores a 86 segundos serán los que se recuperarán. Las horas optimizables estimadas fueron un total de 425 horas.

La estimación de estas horas optimizables es crucial para mejorar la eficiencia operativa y reducir los tiempos de inactividad en el proceso. Al recuperar estas horas, se espera no solo incrementar la productividad sino también optimizar el uso de recursos, minimizando así los costos asociados con el tiempo de inactividad no planificado.

Por otro lado, se calculó previamente la productividad en el punto de descarga CS03. Este cálculo se llevó a cabo utilizando la data proporcionada, obteniendo la sumatoria de tonelaje movido y dividiéndolo por la sumatoria de las horas de acarreo. Así se obtuvo una productividad de 480 t/h. Este valor de productividad es un indicador clave del desempeño del sistema de carguío y acarreo, reflejando la eficiencia con la que se transporta el material.

Para determinar las toneladas que se podrían recuperar en las horas optimizables, se multiplicaron estas horas por la productividad obtenida. De este modo, se calculó un valor de 204 mil toneladas posibles de recuperar. Este cálculo implica una mejora significativa en la capacidad de producción, ya que equivale aproximadamente a tres turnos de producción en ese punto de descarga en un año. Este aumento potencial en la producción resalta la importancia de identificar y recuperar las horas de inactividad.

En la siguiente tabla se muestran todos los resultados descritos en los párrafos anteriores para la estimación de la pérdida. La tabla incluye los valores de *idletime*, las horas optimizables, la productividad del punto de descarga CS03 y las toneladas recuperables.

Esta representación tabular facilita la comprensión de los datos y la relación entre los diferentes factores involucrados en la optimización del proceso.

Tabla 11: Estimación de horas y toneladas optimizables

Estimación de Pérdida	
Intervalo de horas optimizables	
Queue promedio (seg)	95
Queue time objetivo (seg)	86
Horas Optimizables	
425	
Tonelaje Optimizable (Kt)	
Productividad en punto CS03 (t/h)	480
204	

3.3.2. Entendimiento de los datos

El siguiente paso en la metodología es el entendimiento de los datos que serán utilizados para la creación del modelo mediante algoritmos de Machine Learning. Esta etapa se ha dividido en dos partes: Descripción de los Datos y Exploración de los Datos.

En la parte de Descripción de los Datos, se determinará el volumen de los datos y se clasificarán las variables. Este análisis preliminar es crucial para comprender la estructura y el tamaño del conjunto de datos, así como para identificar las características clave de cada variable. Conocer el volumen y la clasificación de los datos permite establecer una base sólida para el análisis posterior.

La Exploración de los Datos implica un análisis más detallado de las variables según su tipo, ya sean numéricas o cualitativas. Para las variables numéricas, se utilizarán técnicas estadísticas descriptivas para entender su distribución, tendencia central y dispersión. En el caso de las variables cualitativas, se analizarán las frecuencias y distribuciones de las categorías, lo cual es esencial para identificar patrones y relaciones dentro de los datos que pueden influir en la creación del modelo de Machine Learning.

3.3.2.1. Descripción de los datos:

En primer lugar, se establecerá con claridad el objetivo del modelo propuesto, el cual se enfoca en predecir si, en el próximo viaje o ciclo de un camión, el tiempo de espera en el punto de descarga excede los 86 segundos. Por lo tanto, este modelo se clasifica como aprendizaje supervisado, dado que se dispone de una variable objetivo; específicamente, se trata de una clasificación debido a la naturaleza binaria del objetivo.

Posteriormente, se detallará el porcentaje de cada clase representada en la variable objetivo. En este contexto, se observa un porcentaje aproximado del 53% para la clase considerada como defectuosa. Este análisis proporciona una visión valiosa para desarrollar un modelo sólido, evitando así el riesgo de un entrenamiento deficiente debido a la falta de

datos en una clase particular.

Finalmente, se presentará una descripción exhaustiva de la cantidad de variables numéricas y categóricas disponibles en el modelo propuesto. Es fundamental distinguir el tipo de cada variable, ya que esto influye significativamente en el análisis subsiguiente y en los tratamientos aplicados a cada una de ellas durante el proceso de modelado. Esta diferenciación facilita la implementación de estrategias efectivas para abordar cada tipo de variable, maximizando así el rendimiento del modelo final.

Tabla 12: Descripción General de las variables del modelo

Objetivo	Determinar si en la siguiente ventana de 2 horas en el turno el promedio de idletime en chancado será mayor a 86 segundos		
Unidad de análisis	Dos horas de cada turno realizado en el año 2016		
Variable dependiente	1	El *idletime* promedio de la siguiente ventana > 86 segundos	52,7%
	0	El idletime promedio de la siguiente ventana < 86 segundos	47.3%
Variables Independientes	28 variables numéricas		
	1 variables categóricas		

3.3.2.2. Análisis Exploratorio de Datos:

Como se mencionó previamente este análisis dependerá de la naturaleza de la variable por lo que será separado en un análisis para variables numéricas y otro para variables categóricas. El análisis exploratorio de datos (AED) incluye la organización, descripción y resumen de los datos. Mediante este, a través de análisis numéricos y representaciones gráficas, se destacan las características que presentan los datos a fin de detectar posibles errores en ellos y, sobre todo, descubrir cuáles son las pautas que los caracterizan. (Escobar Modesto 2013)

a) Análisis de Variables Categóricas

Dentro de las variables categóricas proporcionadas, se llevó a cabo una exploración de datos. Estas variables son fundamentales para entender y predecir los tiempos de espera en el punto de descarga.

En un análisis detallado, se observa que en este punto de descarga solo se utilizaron 110 de los 113 camiones disponibles, lo que sugiere un nivel de utilización del 97%. Además, se identifica que el camión que más descargas realizó fue el CA71, lo que puede ser un indicador importante para comprender los patrones de uso y distribución de los recursos en la operación minera.

En cuanto a los equipos de carguío, se utilizaron 13 de los 14 equipos disponibles, lo que indica una tasa de utilización del 92.8%. El equipo más utilizado fue la pala PA12, lo que

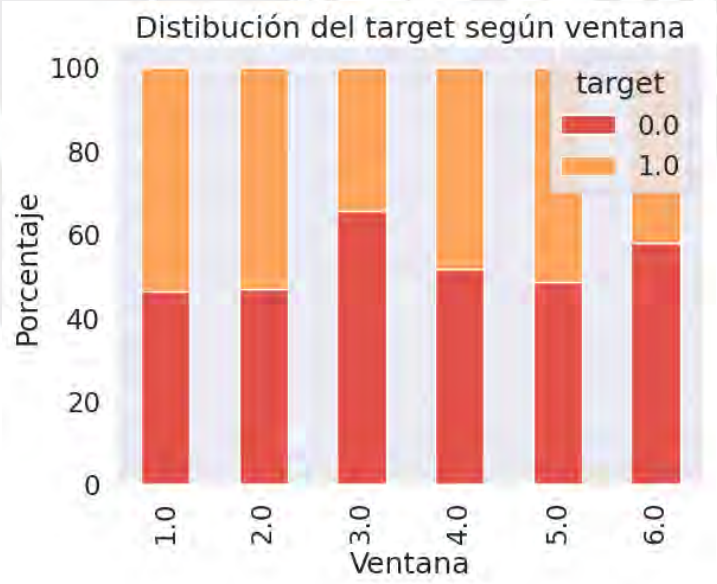
podría sugerir su eficacia o disponibilidad preferida por parte del personal operativo.

Estos hallazgos proporcionan una visión valiosa sobre el rendimiento y la eficiencia en el punto de descarga, así como posibles áreas de mejora en la gestión de recursos y programación de actividades. Un análisis más profundo de estas variables categóricas podría revelar patrones adicionales y oportunidades de mejora a fin de optimizar los procesos de carga y descarga en la operación minera.

Tabla 13: Descripción de las variables truck y excav del modelo

	truck	excav
cantidad	164,656	161,656
únicos	110	13
moda	CA71	PA12

Otro análisis crucial que se realizó fue la variable "seq", que representa la ventana de dos horas dentro de un turno de operación. Esto con la finalidad de entender el comportamiento del target en cada etapa del turno. Para esto, se generó un gráfico de barras apiladas (Gráfica 16).



Gráfica 16: Gráfico de barras apiladas Distribución del target según ventana

En el gráfico se observa que en la ventana 3 se controla mejor la cola en el punto de descarga CS03. Este tipo de insights aportan al modelo para diferenciar qué valor tomará el target según la ventana de tiempo en el turno en que se encuentre.

b) Análisis de Variables Numéricas

Las variables de tipo cuantitativas o numéricas, debido a su naturaleza, requieren otro tipo de gráficos e indicadores para su análisis. En primer lugar, se realizarán histogramas para ver la distribución de las variables y, por otro lado, se realizarán gráficos de cajas para determinar cuáles de estas contienen valores atípicos los cuales deberán ser imputados o eliminados para el desarrollo adecuado del modelo predictivo. Además, se presenta la tabla resumen de los datos estadísticos principales de las variables numéricas.

Variables	cola_ch_promedio	ratio_dumpingtime_NOCH	ratio_dumpingtime_CH	tons_descarga	tons_ch	cola_pala_cargadores_ratio	cola_pala_shovel_ratio
count	4392	4392	4392	4392	4392	4392	4392
mean	92.0	1.9%	0.9%	59664.6	11332.0	0.3%	5.3%
std	67.4	0.6%	0.5%	11978.4	4769.2	0.4%	1.6%
min	0.0	0.0%	0.0%	0.0	0.0	0.0%	0.0%
max	434.5	4.6%	2.5%	308127.9	68243.9	3.0%	12.2%

Tabla 14: Estadísticos principales de las Variables Numéricas

Lo primero que destaca en las tablas es que todas las variables tienen datos, lo que significa que no hay valores vacíos. No obstante, se observa una gran disparidad entre la media y la mediana, lo que sugiere la posible presencia de valores atípicos. Esta suposición será verificada mediante la visualización de gráficos de caja (Tabla 16). Este análisis ayudará a garantizar la integridad y precisión de los datos utilizados en el estudio, lo que a su vez fortalecerá la fiabilidad de los resultados obtenidos.

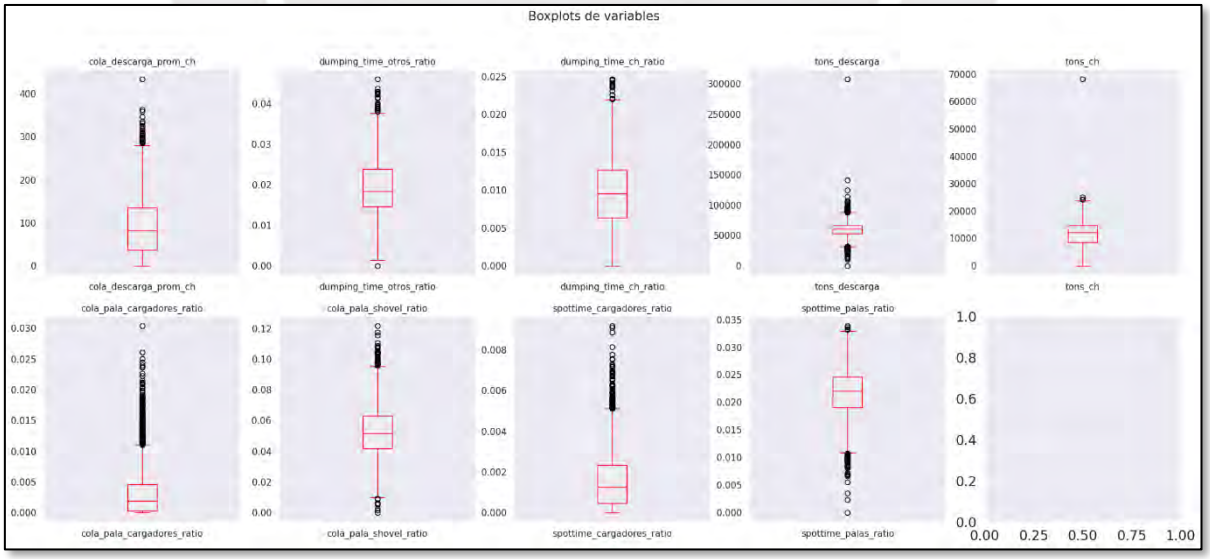
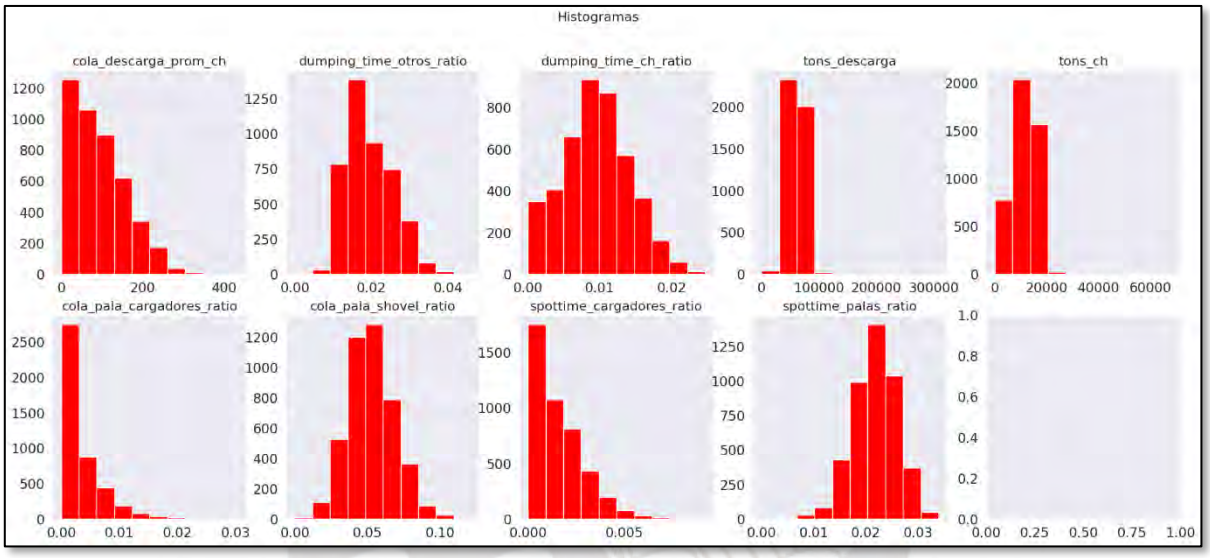


Tabla 15: Gráficos de Caja de variables numéricas

Al utilizar estos gráficos, podemos corroborar la existencia de valores atípicos dentro de nuestros datos, como se sospechaba a partir de los indicadores estadísticos. Además,

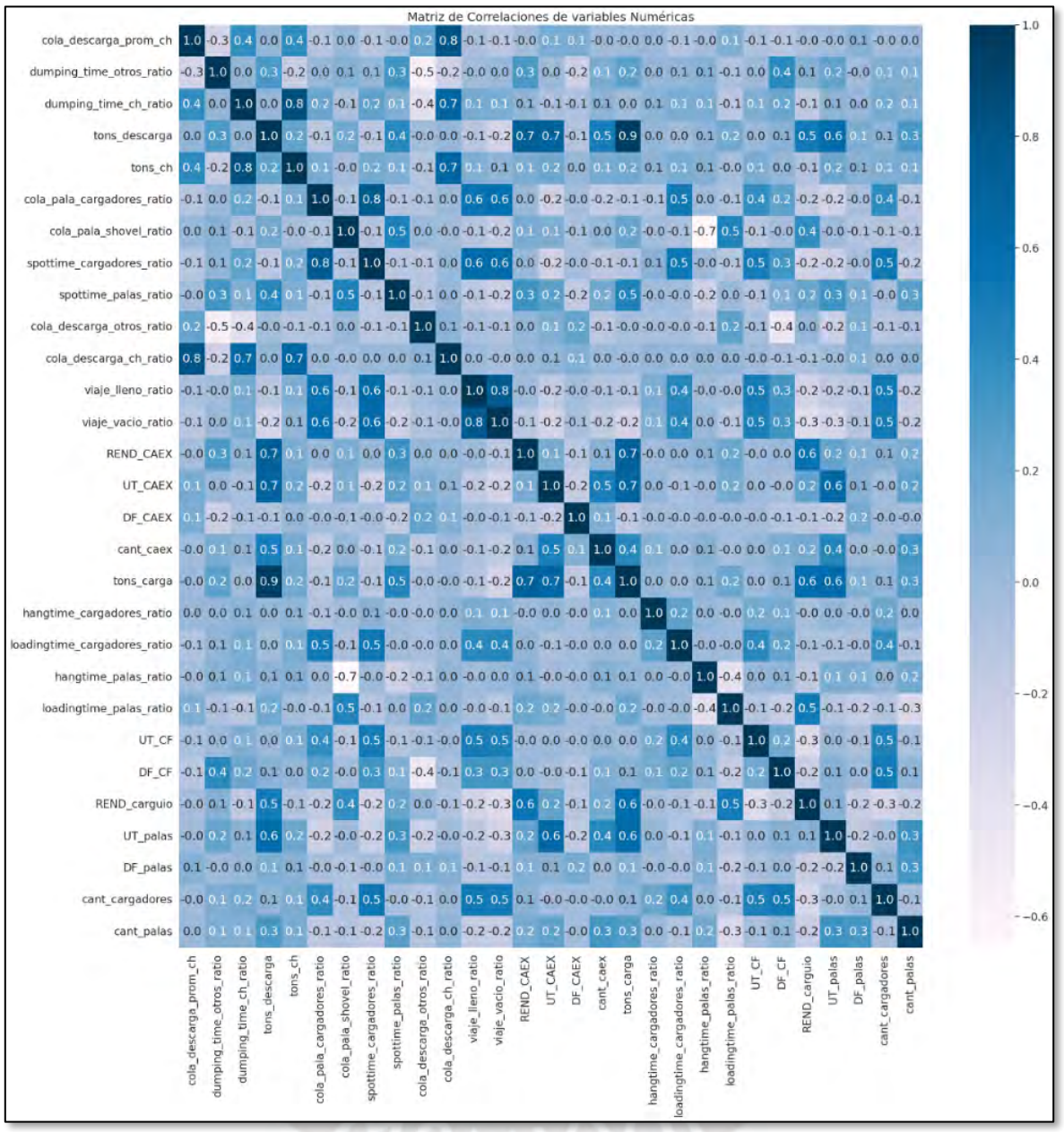
este tipo de representación gráfica nos proporciona una idea de la distribución que siguen las variables. Sin embargo, para obtener una comprensión más detallada de la distribución, se realizarán histogramas de las variables correspondientes (Gráfica 19). Estos histogramas nos permitirán examinar con mayor precisión la forma y la dispersión de los datos, lo que facilitará la identificación de patrones y tendencias importantes en nuestra información. Este enfoque holístico en el análisis de datos nos ayudará a obtener una comprensión más completa y precisa de la naturaleza de nuestras variables, lo que a su vez respaldará el proceso de toma de decisiones en nuestro estudio.



Gráfica 17: Histograma de las variables numéricas del modelo

En estas representaciones gráficas, se observa que algunas variables exhiben distribuciones que son favorables para el desarrollo de modelos predictivos mediante Machine Learning, ya que una distribución conocida facilita su predictibilidad. Sin embargo, otras variables muestran distribuciones en las que resulta difícil detectar patrones evidentes, pero se espera abordar este desafío mediante ingeniería de variables y transformaciones adecuadas.

Un análisis crucial en variables cuantitativas es el de correlaciones entre ellas, ya que una relación notable entre este tipo de variables puede conducir a un sobre entrenamiento del modelo. Para llevar a cabo este análisis de manera efectiva, se empleará la matriz de correlaciones (Gráfica 20). El coeficiente de correlación de Pearson será utilizado para determinar la fuerza y la dirección de la relación entre las variables numéricas, dado que es el más ampliamente utilizado para este propósito. Este enfoque nos permitirá evaluar la interdependencia entre las variables y ajustar nuestro modelo de manera apropiada para evitar problemas de sobre entrenamiento y mejorar la precisión de nuestras predicciones. (Mendenhall, Beaver, and Beaver 2010).



Gráfica 18: Matriz de correlación de Pearson de las variables numéricas

3.3.3. Preparación de los datos

Para el desarrollo de modelos predictivos utilizando herramientas computacionales como el Machine Learning, es fundamental cumplir con ciertos requisitos previos. Primero, es imprescindible que todos los datos estén en formato numérico, ya que los algoritmos de Machine Learning requieren este tipo de datos para procesar y generar resultados precisos. Además, se deben llevar a cabo diversas transformaciones y tratamientos de los datos para asegurar que los inputs al modelo estén optimizados y permitan obtener resultados positivos. Estos tratamientos incluyen la normalización, estandarización, imputación de valores

faltantes, y la codificación de variables categóricas, entre otros.

En esta sección, se detallarán los pasos necesarios para lograr los objetivos previamente descritos. Se explicará de manera exhaustiva cómo se llevarán a cabo las transformaciones y tratamientos de los datos, así como los métodos utilizados para la clasificación y selección de las variables. El objetivo es establecer una metodología clara y rigurosa que garantice la calidad y efectividad de los modelos predictivos desarrollados.

3.3.3.1. Codificación de variables Categóricas y Transformación de atípicos

Esta etapa se dividirá en dos fases. La división se realizará en función del tipo de datos, ya que las variables categóricas requieren un tratamiento diferente al de las variables numéricas. En el caso de las variables categóricas, estas necesitan ser codificadas en números, un proceso que puede involucrar técnicas como la codificación ordinal, la codificación de etiquetas o la creación de variables dummy.

Por otro lado, las variables cuantitativas requieren un tratamiento especial para manejar los valores atípicos. Este proceso puede incluir la detección y eliminación de outliers, la imputación de valores extremos, o la transformación de datos para reducir su impacto en el modelo. Estas acciones son esenciales para asegurar que los datos sean coherentes y adecuados para el análisis predictivo, minimizando el riesgo de sesgos y mejorando la precisión del modelo.

Ambas fases son cruciales y complementarias para el preprocesamiento de datos, ya que garantizan que tanto las variables categóricas como las cuantitativas se ajusten adecuadamente a los requisitos del modelo de Machine Learning, optimizando así su rendimiento y precisión en la etapa de predicción.

a) Codificación de Variables Categóricas

Por lo general, cualquier conjunto de datos estructurados incluye varias columnas, que son una combinación de variables numéricas y categóricas. Las máquinas solo pueden entender números; no pueden comprender texto. Esto es aplicable también a los algoritmos de Machine Learning, que requieren datos numéricos para su procesamiento. Por esta razón, es necesario convertir las columnas categóricas en columnas numéricas, un proceso conocido como codificación categórica (SETHI 2020).

Las variables categóricas presentes en este trabajo son de tipo ordinales, ya que son etiquetas ordenadas asignadas a diferentes clases. Por lo tanto, el tipo de codificación más adecuado para este tipo de variable es la codificación por etiquetas (Label Encoding). En esta técnica, a cada etiqueta se le asigna un número entero único basado en el orden

alfabético (SETHI 2020).

A modo de ejemplo, se presentarán los datos antes y después de ser codificados, para ilustrar cómo han variado debido al tratamiento necesario para el procesamiento posterior. Estos se muestran en las siguientes tablas:

Tabla 16: Variables categóricas antes de la codificación

Var	seq
0	1
1	2
2	3
3	4
4	5

Tabla 17: Variables categóricas después de la codificación

Var	hos_x
0	**0**
1	**1**
2	**2**
3	**3**
4	**4**

b) Tratamiento de atípicos en Variables Numéricas

En el análisis exploratorio de los datos, se detectaron valores atípicos en las variables numéricas mediante el uso de diagramas de cajas. Estos valores atípicos no son deseables en los modelos predictivos, ya que pueden provocar un aprendizaje incorrecto por parte del algoritmo. Por esta razón, se implementó una estrategia sugerida por la empresa que nos proporcionó los datos.

La metodología utilizada para tratar los valores atípicos consistió en reemplazar los valores superiores al percentil 97 por el valor correspondiente a este percentil. Esta técnica permite reducir la cantidad de outliers presentes en los datos

3.3.3.2. Ingeniería de Variables:

Esta etapa consiste en crear nuevas variables a partir de las ya existentes en los datos. De esta manera, se pueden solucionar problemas como la correlación entre variables o la falta de escalamiento de las mismas (mediante el uso de logaritmos).

Los algoritmos de aprendizaje automático utilizan ciertos datos de entrada para producir resultados. Sin embargo, con bastante frecuencia, los datos proporcionados pueden no ser suficientes para diseñar un modelo de aprendizaje automático efectivo. Aquí es donde

entra en juego el poder de la Ingeniería de Variables (Analytics Vidhya 2020). Los principales objetivos de esta son:

- Preparar un conjunto de datos de entrada adecuado, compatible con los requisitos del algoritmo de Machine Learning.
- Mejorar el rendimiento de los modelos de Machine Learning.

En el presente trabajo de tesis, se realizaron tres transformaciones. En primer lugar, se crearon nuevas variables a partir de variables existentes. A continuación, se muestra los tipos de cálculos que se realizaron:

(1) Rendimiento: Movimiento por hora operativa

$$\frac{\boldsymbol{Tons}}{\boldsymbol{Tiempo\ operativo}} = \boldsymbol{Rendimiento}$$

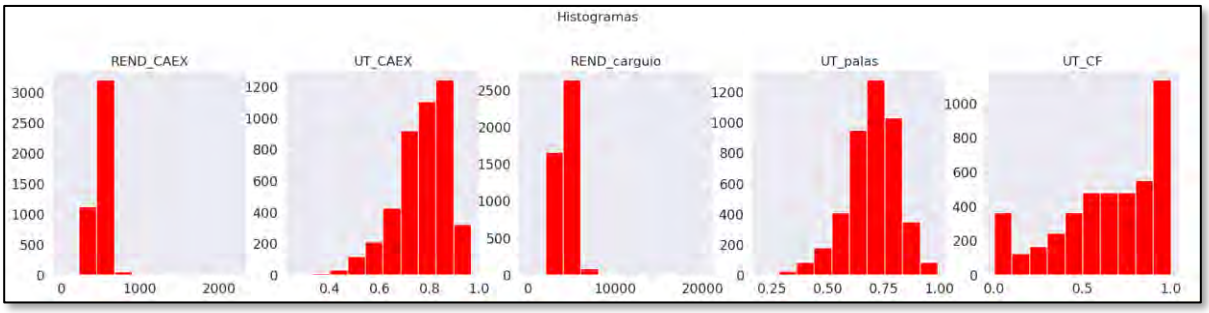
(2) Utilización: Porcentaje que representa el tiempo en el que un determinado equipo se encuentra en uso del tiempo total disponible.

$$\frac{\boldsymbol{Tiempo\ Operativo}}{\boldsymbol{Tiempo\ Disponible}} = \boldsymbol{Utilización}$$

(3) Ratios del Tiempo de Ciclo: Estos representan la fracción del tiempo operativo que ocupa cada tiempo de ciclo del proceso.

Es importante resaltar que los datos a los que aplica la transformación logarítmica deben tener solo valores positivos.

A continuación, se muestran los histogramas de las nuevas variables creadas para tener una idea de la distribución de estas.



Gráfica 19: Distribución de nuevas variables creadas

3.3.3.3. Selección de Variables:

La selección de variables es un procedimiento crítico en el aprendizaje automático, ya que determina qué variables se utilizarán para alimentar el modelo y realizar predicciones, ya sea para regresión o clasificación. Es fundamental evitar la inclusión de variables irrelevantes, ya que estas pueden introducir ruido innecesario en los datos y reducir la precisión del modelo. Para abordar este problema, se suelen emplear métodos estadísticos como ANOVA o la prueba de chi-cuadrado, los cuales evalúan la relación entre cada variable predictiva y la variable objetivo (Wong 2020).

Los resultados de estos métodos se presentarán en una tabla resumen que indicará qué variables han sido seleccionadas siguiendo una metodología específica. En este trabajo, se han utilizado tres métodos reconocidos: ANOVA, WOES y Random Forest. Estos métodos proporcionarán una visión integral de las variables más relevantes para el modelo, lo que contribuirá a mejorar su desempeño predictivo y su capacidad para generalizar patrones en los datos.

	Driver's Selection Method				
Variable	Random Forest		WOES		Anova
cola_descarga_prom_ch	0.0692	1	0.426747	1	yes
cola_descarga_ch_ratio	0.0589	3	0.353913	2	yes
shiftindex	0.0606	2	0.227521	3	yes
dumping_time_otros_ratio	0.0412	4	0.154537	4	yes
cola_descarga_otros_ratio	0.0411	5	0.122975	5	yes
DF_CF	0.0196	28	0.121978	6	yes
tons_ch	0.0389	7	0.114807	7	yes
loadingtime_palas_ratio	0.0348	8	0.053326	8	yes
cant_caex	0.0248	26	0.039135	9	yes
dumping_time_ch_ratio	0.0393	6	0.030516	10	yes
UT_CF	0.0270	22	0.024589	11	yes
DF_CAEX	0.0330	11	0.023795	12	yes
DF_palas	0.0267	23	0.019127	13	yes
viaje_vacio_ratio	0.0279	21	0.019116	14	yes
loadingtime_cargadores_ratio	0.0283	20	0.011250	15	yes
UT_CAEX	0.0310	17	0.010289	16	yes
tons_carga	0.0338	10	0.007535	17	yes
REND_carguio	0.0329	12	0.007374	18	
hangtime_palas_ratio	0.0345	9	0.006613	19	yes
cola_pala_cargadores_ratio	0.0243	27	0.005647	20	
spottime_cargadores_ratio	0.0260	25	0.004304	21	yes
viaje_lleno_ratio	0.0288	19	0.003974	22	
tons_descarga	0.0323	14	0.003526	23	
UT_palas	0.0300	18	0.002903	24	
cant_cargadores	0.0053	30	0.002684	25	
hangtime_cargadores_ratio	0.0263	24	0.001763	26	
cola_pala_shovel_ratio	0.0325	13	0.001648	27	
spottime_palas_ratio	0.0322	15	0.001173	28	
REND_CAEX	0.0311	16	0.000241	29	
cant_palas	0.0094	29	0.000134	30	

Gráfica 20: Resultados de la Selección de Variables

3.3.4. Modelamiento

El proceso de creación de un modelo predictivo consta de tres pasos importantes: la partición, el entrenamiento y la evaluación. En este primer trabajo se realizará una corrida de los algoritmos. Esta corrida se efectuará con todas las variables, es decir, sin hacer uso de la selección de variables. En ese sentido, este primer modelo servirá como línea base para posteriores modelos ajustados. Los algoritmos que se usarán son:

(1) Regresión Logística
(2) Random Forest para la Clasificación
(3) Adabbost

3.3.4.1. Partición Muestral de los datos

En este primer paso, los datos se separan en dos grupos: el primero para entrenar el modelo y el segundo para hacer pruebas del modelo entrenado, es decir, para validarlo. Por ello, se les asigna el nombre de datos de entrenamiento y datos de evaluación. Los porcentajes y volúmenes para cada uno de ellos se muestran en la siguiente tabla.

Estos porcentajes utilizados son los recomendados para un modelo de clasificación de aprendizaje automático.

Data Set	Número de Registros	Porcentaje (%)
Data de Entrenamiento	2,925	66.6%
Data de Evaluación	1,467	33.4%
Total	4,392	100%

Tabla 18: Partición muestral de los datos

3.3.4.2. Entrenamiento, Resultados y Evaluación de Resultados

Luego de realizar la partición de los datos, se procede al entrenamiento de los algoritmos mencionados previamente. Los resultados que se muestran incluyen la matriz de confusión, ya que este es un método muy utilizado para la evaluación de modelos de clasificación. Como se aclaró previamente, estos primeros resultados provienen de una primera corrida con todas las variables disponibles. Posteriormente, se pueden desarrollar otros modelos seleccionando variables específicas, utilizando diferentes variables o con una mayor cantidad de datos.

a) Regresión Logística

En la Tabla 22 se muestran los resultados de la regresión logística: La matriz de

confusión y los indicadores obtenidos a partir de esta.

TRAIN

	CONFUSION MATRIX	
PREDICTION	1091	451
	562	821
	TRUE VALUE	

Accuracy	65.37%
Precision	64.54%
Recall	59.36%

TEST

	CONFUSION MATRIX	
PREDICTION	532	241
	274	420
	TRUE VALUE	

Accuracy	64.89%
Precision	63.54%
Recall	60.52%

Tabla 19: Indicadores técnicos del modelo de Regresión Logística

b) Random Forest para la Clasificación

En la Tabla 20se muestran los resultados de la regresión logística: La matriz de confusión y los indicadores obtenidos a partir de esta.

TRAIN

	CONFUSION MATRIX	
PREDICTION	1121	421
	519	864
	TRUE VALUE	

Accuracy	67.86%
Precision	67.24%
Recall	62.47%

TEST

	CONFUSION MATRIX	
PREDICTION	533	240
	272	416
	TRUE VALUE	

Accuracy	64.96%
Precision	63.41%
Recall	60.47%

Tabla 20: Indicadores Técnicos del modelo de Random Forest

c) Adaboost

En la Tabla 21 se muestran los resultados de la regresión logística: La matriz de confusión y los indicadores obtenidos a partir de esta.

TRAIN

	CONFUSION MATRIX	
PREDICTION	1220	322
	346	1037
	TRUE VALUE	

Accuracy	77.16%
Precision	76.31%
Recall	74.98%

TEST

	CONFUSION MATRIX	
PREDICTION	517	256
	282	412
	TRUE VALUE	

Accuracy	63.33%
Precision	61.68%
Recall	59.37%

Tabla 21: Indicadores Técnicos del modelo de Adaboost

4. RESULTADOS

Con el análisis del gráfico de Pareto se llega a determinar que las variables importantes que pueden optimizarse son el Idletime y Hangtime. Esta oportunidad de mejora se basa en que se tiene gran cantidad de horas en estas pérdidas operacionales y que se cuenta con la data disponible para hacer el análisis y la posterior optimización.

Analizando el tiempo de inactividad agrupado por puntos de descarga, se detecta que el punto de descarga de mineral es donde el tiempo de inactividad no está bajo control. Esto se demuestra ya que tiene una variabilidad elevada. Por otro lado, los otros puntos de descarga no presentaban este problema.

Controlando la variabilidad del tiempo de inactividad, se mejorará la productividad (tonelaje) porque se va a ahorrar tiempo efectivo. En ese sentido, al mejorar la productividad en el punto de descarga del mineral reduciría los costos de la empresa minera y esto tendría un impacto positivo en las utilidades de la empresa.

De los resultados obtenidos en las tablas de evaluación de los modelos se puede ver que el modelo que dio el peor performance fue el Adaboost. Esto se debe, probablemente, al sobre entrenamiento del modelo con este algoritmo ya que se ven buenos resultados en el entrenamiento, pero no durante la evaluación. Esta es una característica de los modelos avanzados como el Adaboost, la cual se podría optimizar posteriormente.

Se selecciona para el presente modelo el algoritmo de regresión logística ya que logra buenos resultados en los indicadores, tanto en entrenamiento como en evaluación. De esta manera se obtiene un modelo ni sobre ni sub entrenado con un accuracy de 64.96%; es decir, se aciertan en, aproximadamente, 8 de 12 predicciones durante el día.

ET LUX IN TENEBRIS LUCET
MCMXVII

5. CONCLUSIONES

- Es posible desarrollar y evaluar un modelo predictivo utilizando Machine Learning para optimizar una etapa de un proceso minero.

- Todo análisis de oportunidad de mejora utilizando Machine Learning debe realizarse con suficiente data disponible.

- Es necesario determinar las variables importantes con las que se alimentara el algoritmo, así optimizar el rendimiento del modelo de predicción.

- Existen actividades en el proceso minero que presentan una variabilidad elevada, en nuestro caso de estudio esta fue la cola en el punto de descarga de mineral.

- No todos los modelos brindan un óptimo resultado. En nuestro caso, el modelo Adaboost fue el que brindó un performance nada satisfactorio ya que este sobreentrenó el modelo.

- El modelo de algoritmo de regresión logística es el que brindo los mejores resultados en los indicadores con una exactitud de 64.96%. Obtuvo buenos resultados en los datos de entrenamiento y de evaluación.

- La preparación adecuada de los datos es indispensable para obtener resultados positivos en los modelos de aprendizaje automático. De qué tan bien se haga esta etapa de la metodología depende el performance de los modelos resultantes.

ET LUX IN TENEBRIS LUCET
MCMXVII

6. RECOMENDACIONES

- Antes de realizar cualquier tipo de modelo se debe hacer un análisis profundo y exhaustivo del negocio minero para poder detectar mejoras que tengan un impacto significativo sobre este.
- Existe una gran variedad de algoritmos de aprendizaje automático, en este sentido se debe ensayar con distintos algoritmos para encontrar cual de todos es el más adecuado para el caso que se venga estudiando.
- Realizar una línea base del modelo permitirá hacer comparaciones con posteriores resultados y, de esta manera, darse cuenta de la importancia de los nuevos modelos creados; es decir, si se generan mejoras o no.
- Los modelos predictivos de aprendizaje automáticos deben ser utilizados como una herramienta para la toma de decisiones. No debe reemplazar la decisión de las personas encargadas, sino estas deben utilizar los resultados de los modelos para tomar decisiones que puedan generar impactos positivos en la actividad minera, en este caso reduciendo perdidas de horas operacionales.

7. REFERENCIAS

Aggarwal, Charu C. 2016. *Recommender Systems*. Vol. 40.

Analytics Vidhya. 2020. “7 Feature Engineering Techniques | Feature Engineering for ML.” Retrieved June 24, 2021 (https://www.analyticsvidhya.com/blog/2020/10/7-feature-engineering-techniques-machine-learning/).

Baek, Jieun, and Yosoon Choi. 2019. “Simulation of Truck Haulage Operations in an Underground Mine Using Big Data from an ICT-Based Mine Safety Management System.” *Applied Sciences (Switzerland)* 9(13).

Baek, Jieun, and Yosoon Choi. 2020. “Deep Neural Network for Predicting Ore Production by Truck-Haulage Systems in Open-Pit Mines.” *Applied Sciences (Switzerland)* 10(5).

Barreto Taipe, Lides. 2017. “Optimizacion Del Numero de Camiones 785C CAT y Cargador Frontal 992K CAT Mediante El Match Factor En La Ruta Mineral – Stock Pile Antapaccay – Chancadora Tintaya San Martin Contratistas Generales S.A.”

Barroso Salgado, Javier. 2019. “Modelo Predictivo Basado En Machine Learning de Órdenes de Trabajo Riesgosas Para Mantenimiento de Equipos Mineros.”

Bonzi, José Ignacio. 2016. “Memoria Para Optar Al Título de Ingeniero Civil de Minas: Propuestas de Mejora de La Utilización Efectiva En Base a Disponibilidad de La Flota de Carguío y Transporte En Minera Los Pelambres.” 122.

Bustamante Chávez, José Eder. 2018. “Optimización de la productividad de los equipos de carguío y acarreo en gold fields la cima s.a mediante la disminución de las demoras operativas más significativas.” *Photosynthetica* 2(1):1–13.

Chaowasakoo, Patarawan, Heikki Seppälä, Heikki Koivo, and Quan Zhou. 2017. “Digitalization of Mine Operations: Scenarios to Benefit in Real-Time Truck Dispatching.” *International Journal of Mining Science and Technology* 27(2):229–36.

Escobar Modesto. 2013. *Análisis Gráfico y Exploratorio de Los Datos.* Vol. 53.

Gonzales, Ligdi. 2020. *Machine Learning Con Python Aprendizaje Supervisado*.

Hunt, John. 2019. *A Beginners Guide to Python 3 Programming*.

IBM. 2020a. "¿Qué Es Machine Learning? - Perú | IBM." Retrieved November 24, 2020 (https://www.ibm.com/pe-es/analytics/machine-learning).

IBM. 2020b. "Conceptos Básicos de Ayuda de CRISP-DM." Retrieved December 7, 2020 (https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_crispdm_ddita/clementine/crisp_help/crisp_overview.html).

Igual, Laura, and Santi Seguí. 2017. *Introduction to Data Science*.

Infante, Freddy Calua. 2019. "Propuesta de Minimización de Tiempos Improductivos Para Una Mayor Producción En Carguío y Acarreo En CIA. Minera Coimolache S.A."

Kubat, Miroslav. 2017. *An Introduction to Machine Learning*.

Medina, Fernando, and Marco Galván. 2007. *Imputación de Datos: Teoría y Práctica*. Vol. 4.

Mendenhall, William, Robert Beaver, and Barbara Beaver. 2010. *Introducción a La Probabilidad y Estadística*.

Pavlov, Yu L. 2019. "Random Forests." *Random Forests* 1–122.

Sahoo, Kabita, Abhaya Kumar Samal, Jitendra Pramanik, and Subhendu Kumar Pani. 2019. "Exploratory Data Analysis Using Python." *International Journal of Innovative Technology and Exploring Engineering* 8(12):4727–35.

Sethi, Alakh. 2020. "Categorical Encoding | One Hot Encoding vs Label Encoding." Retrieved June 24, 2021 (https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/).

Skiena, Steven. 2017. *Data Science Design*. Cham: Springer International Publishing.

Vidal, Manuel. 2010. *Estudio del cálculo de flota de camiones para una operación*

minera a cielo abierto.

Wong, Jason. 2020. “Feature Selection With BorutaPy. This Post Will Serve as a Tutorial On… | by Jason Wong | Towards Data Science.” Retrieved June 24, 2021 (https://towardsdatascience.com/feature-selection-with-borutapy-f0ea84c9366).