TI3001C Analítica de datos y herramientas de inteligencia artificial I
LU, JU de 15:00 a 21:00 / MA, MI, VI de 17:00 a 21:00
Grupo 104

EVIDENCIA 1. ANÁLISIS ESTADÍSTICO DE LA BASE DE DATOS

Equipo 3: Chameleon

Nayeli Peña Martínez A01368516

Karla Sánchez Del Ángel A01198184

Ania Acosta Castro A01736506

Alejandro Correa Amaya A01028507

Profesor:

Oscar Daniel Ortíz Esquivel

Monterrey, Nuevo León

17 de octubre de 2024

Introducción, objetivo y contexto de la empresa

Ternium, es una destacada empresa siderúrgica en las Américas, cuenta con 18 centros de producción y 51 centros de servicio y distribución en diversos países de América Latina y Estados Unidos. Con una plantilla de 20,500 empleados, la compañía gestiona todo el proceso, desde la extracción de hierro hasta la producción y distribución de productos de acero.

Misión: "Crear valor con nuestros clientes, mejorando la competitividad y productividad conjunta" (Vedoya, 2022).

Visión: "Ser la empresa siderúrgica líder de América, comprometida con el desarrollo de sus clientes, a la vanguardia en parámetros industriales y destacada por la excelencia de sus recursos humanos" (Vedoya, 2022).

Objetivo de la empresa: Ofrecer valor a los stakeholders de la empresa mientras la empresa se posiciona como líder en América Latina y es competitiva en el mercado. Además, se enfocan en brindar productos y servicios de manera sustentable.

Objetivos del proyecto

El objetivo general del proyecto de analítica de datos es incrementar la productividad del proceso de laminación en caliente 4 de Pesquería mediante alguno de los siguientes objetivos secundarios:

- Identificar qué influye en que la productividad no sea la esperada, buscar soluciones para mitigar los cuellos de botella y buscar que los cuellos de botella sucedan más en la fase FCE que en la fase FM.
- Identificar las causas raíz de las interrupciones y micro demoras que ralentizan la producción.
- Brindar insights accionables para hacer que el proceso de producción sea más eficiente y fuerte frente a variaciones en las condiciones operativas.

Set de Datos

Para esta etapa del proyecto se utilizaron dos bases de datos las cuales se describen a continuación:

1. Exportación Pacing 2024: Registro del paso de los planchones a través de toda la línea.
2. SGL_Demoras_Caliente 4 PES: Interrupciones y paradas generadas en la línea; incluye el tipo de demora, concepto, subconcepto, comentario y tiempo asociado. La base de datos original contiene un total de 3,377 registros y 18 columnas. En su mayoría, las variables de esta base de datos son de tipo categóricas nominales. La distinción entre interrupciones y paradas es la siguiente:
    - Interrupciones: No programadas, pueden ser operativas o no operativas
    - Paradas: Programadas

Situación Actual

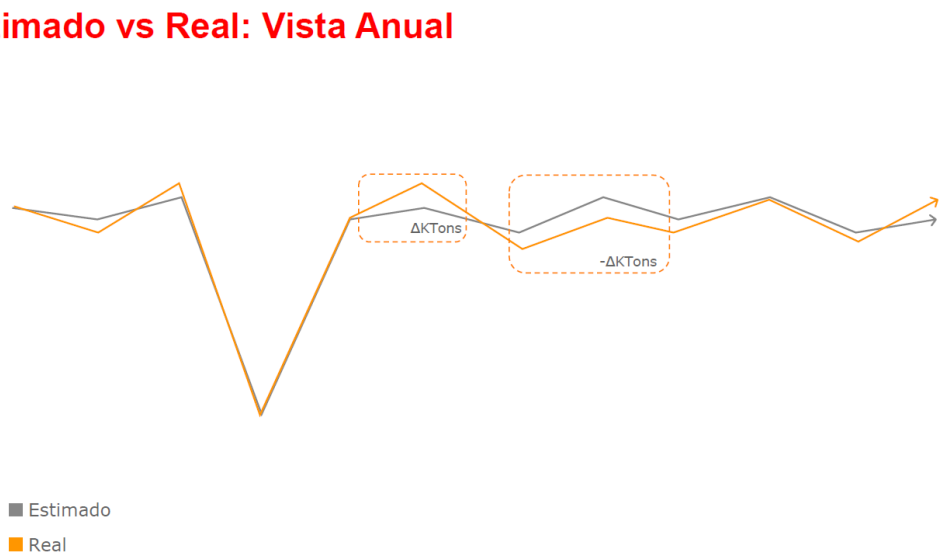

Figura 1. *Línea de Tiempo de Productividad Esperada vs Real.*

La figura 1 muestra las diferencias entre la productividad real y la esperada de Ternium. Esas diferencias enmarcadas con cuadros de color amarillo son las que ellos buscan atacar mediante la analítica de datos para encontrar causas raíz y/o patrones que las expliquen y así poder simular escenarios de posibles soluciones.

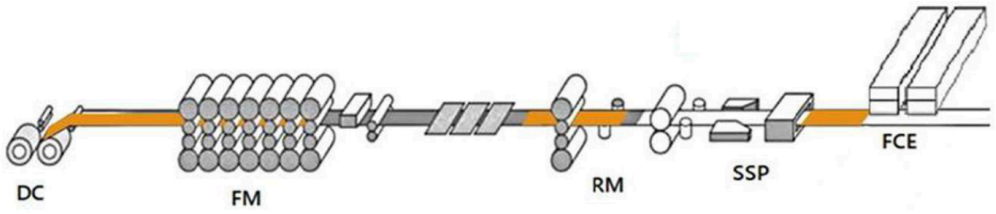

Figura 2. *Proceso de Laminación en caliente del molino caliente 4 de Pesquería.*

El proceso consta de 5 fases en total. Tal y como se mencionó anteriormente, la mayor cantidad de cuellos de botella se encuentran actualmente en el horno y el FM "Finishing Mill". Según Ternium, es normal que un proceso de laminación cuente con cuellos de botella, pero que ellos buscan que estas ocurran más dentro del horno ya que las pueden controlar mejor. Es importante considerar que hay 5 planchones en un mismo horno, cada uno perteneciente a una de las 3 familias de calentamiento y rangos de tolerancia. Esto permite los agrupamientos de anchos y espesores.
Dicho esto, en esta etapa decidimos separar por fases el análisis, y enfocarnos en estudiar la fase del FM. Esto facilitará la identificación de causas raíz y el resultado esperado es la mitigación de las demoras en esta fase, para que la mayoría se encuentren en el horno.

Exploración de los datos

A continuación se explica el análisis exploratorio de los datos de las bases de datos mencionadas previamente.

Base Datos SGL Demoras Caliente 4 PES

Dimensiones: 3,377 filas x 18 columnas
Descripción:

```
Data columns (total 18 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Fecha                3377 non-null   object
 1   DAY                  3377 non-null   int64
 2   MONTH                3377 non-null   object
 3   YEAR                 3377 non-null   int64
 4   Fecha Inicio         3377 non-null   object
 5   Fecha Fin            3377 non-null   object
 6   Escuadra             3377 non-null   object
 7   Duracion             3377 non-null   float64
 8   Tipo                 3377 non-null   object
 9   Rubro                3377 non-null   object
 10  Subrubro             3377 non-null   object
 11  Concepto             3377 non-null   object
 12  Subconcepto          3377 non-null   object
 13  Observaciones        3366 non-null   object
 14  Ub. Tecnica Nivel 1  3377 non-null   object
 15  Ub. Tecnica Nivel 2  3377 non-null   object
 16  Equipo               3377 non-null   object
 17  Rollo                3377 non-null   object
dtypes: float64(1), int64(2), object(15)
```

Figura 3. *Descripción de columnas y sus tipos.*

Distribuciones estadísticas

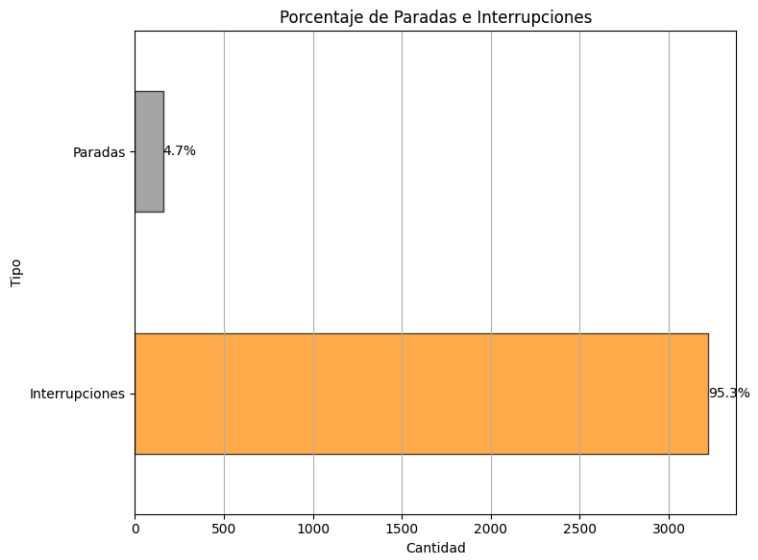

Figura 4. *Distribución de Paradas e Interrupciones.*

*El **95.3%** de los registros corresponden a **interrupciones** o demoras no programadas. El **4.7%** restante de los registros en demoras fueron **programados**.*

Medidas de tendencia central

El tiempo promedio de paradas es de 109.10 minutos, la mediana es de 21.6 minutos y la moda de 360 minutos, equivalente a 6 horas. El tiempo promedio de interrupciones es de 9.23 minutos, la mediana es de 4.1 minutos y la moda de 9.5 minutos.

Del 01 de mayo al 21 de agosto el total de horas de paradas fue de 287.31, mientras que el total de horas de interrupciones fue un total de 495.19. En aproximadamente 3 meses el total de horas 'no productivas' por demoras fue de 782.50.

Medidas de dispersión

La desviación estándar de paradas es de 153.88 minutos.
La desviación estándar de interrupciones es de 19.77 minutos

Base de Datos Exportación Pacing 2024

Dimensiones: 56,798 filas x 113 columnas
Descripción: Paso de planchones en MC4, desde que **entran al horno hasta la salida** después de ser enrollados, incluye dimensiones de planchones. Una de las variables de interés para el análisis es 'Delay Time' ya que cuenta con el tiempo de retraso por cada planchón.

Distribuciones Estadísticas

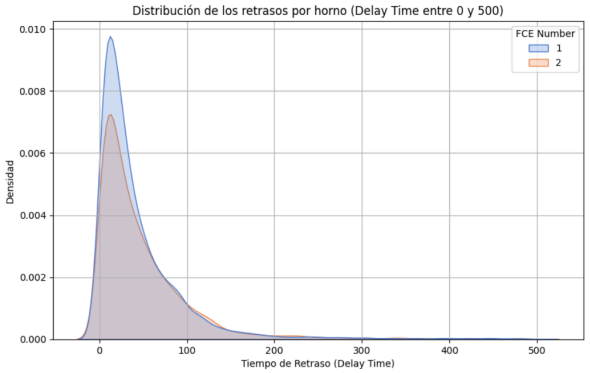

***Figura 5.** Densidad de valores 'Delay Time'*

En cuanto a los tiempos de retraso, se observa como la mayoría de los datos se encuentran en el intervalo [0, 100] segundos, sin embargo, el df cuenta con valores que superan los 500 segundos, por lo que se puede estar tratando de outliers.

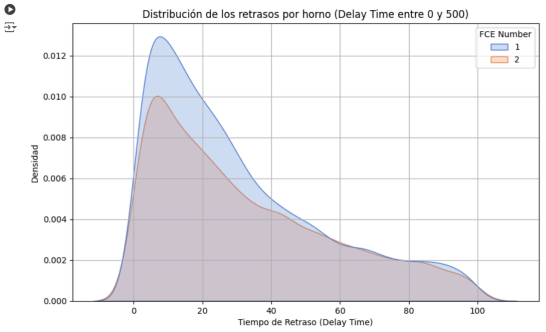

Figura 6. *Densidad de valores 'Delay Time' filtrado de [0,100]*

Se observa una distribución normal con un sesgo hacia la izquierda.

Figura 7. Gráficos *de densidad para las dimensiones de los planchones.*

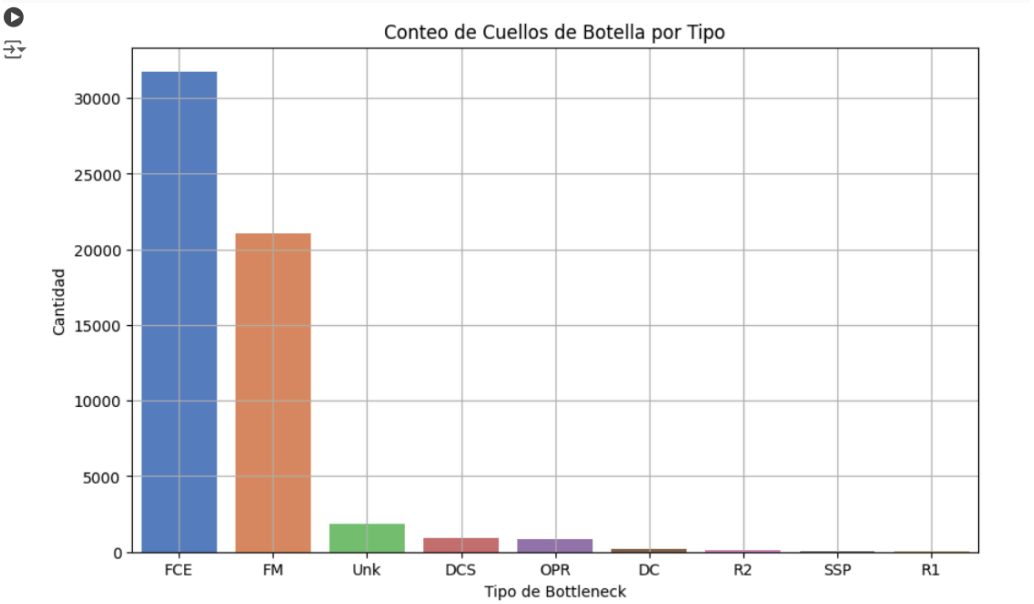

Figura 8. *Distribución de lugares de cuellos de botella.*

Medidas de Tendencia Central

FCE Number	Mean	Median	Std
1	60.53	13	496.01

2	65.44	7	811

Cómo se aprecia en la tabla, ambas columnas presentan datos que alteran el correcto entendimiento de patrones dentro de Delays. La desviación estándar es demasiado alta, en comparación con la media y la mediana. Habrá que identificar los outliers para obtener resultados más verídicos. Por ahora, la media de Delay es de 60.5 y 65.4 para los hornos 1 y 2 respectivamente.

Preprocesamiento & Limpieza

Se unieron las dos bases de datos mencionadas con ayuda del identificador de cada planchón y posteriormente se limpió de la siguiente manera:

1. Eliminación de columnas por irrelevancia
2. Imputación de NaNs y ceros con media y medianas
3. Eliminación de duplicados
4. Agrupación de familias por familia de acero
5. Tratamiento de outliers por x1.5 veces el IQR

4. Modelación:

A continuación se presentan los 4 modelos utilizados en esta etapa y sus respectivas interpretaciones. Para los modelos de clasificación se utilizó como variable independiente la media de 'FM Thread Time',y para regresión simplemente los valores de dicha columna.

4.1 Regresión Lineal Múltiple

Para el modelo de regresión múltiple, como primer paso calculamos la matriz de correlación con la variable `FM Thread Time.` Esta variable es de nuestro interés dado que en esta entrega nos enfocaremos en la fase FM (finishing mill).

```
FM Thread Time              1.000000
1st Accel Time              0.704730
Coil Thickness              0.692724
FM Travel Pacing Time       0.575000
STD Interval Time           0.467072
Slab Width                  0.368901
Taper_Head_Width_Meas       0.362130
Taper_Tail_Width_Meas       0.362009
STD Cycle Time              0.298635
SSP use                     0.271344
Slab Weight                 0.214291
Slab Thickness              0.195998
Slab Lenght                -0.212053
FDT                        -0.350008
2nd Accel Time             -0.436825
steel_type                 -0.443852
R2 Pacing Time             -0.468276
FM Final Speed             -0.773652
FM Thread Speed            -0.876033
```

Figura 9. *Correlación de variables*

Estos resultados nos permitieron seleccionar las variables más relevantes para nuestro modelo de regresión las cuales son:

```
X = data[['FM Thread Speed', 'FM Final Speed', '1st Accel Time', 'FM Travel Pacing Time', 'R2 Pacing Time', 'steel_type', 'Slab Width', 'Taper_Head_Width_Meas', 'Taper_Tail_Width_Meas', 'Slab Weight', 'Slab Lenght', 'Slab Thickness']]
```

Así mismo, se destaca que la selección de variables con mayor coeficiente de correlación corresponde a aquellas relacionadas con la fase FM o previas a la misma.

El modelo tiene una R2 de 0.81 lo que indica un buen ajuste del modelo a los datos. El modelo explica el 81% de variabilidad de la variable dependiente. El MSE es de 2.67. Un error cuadrático medio de este valor nos indica que el modelo tiene un error de de 2.67 segundos para predecir el tiempo de un planchón en FM.

4.2 Regresión Logística

```
X = data[['FM Thread Speed', 'FM Final Speed', '1st Accel Time', 'FM Travel Pacing Time',
      'R2 Pacing Time', 'steel_type', 'Slab Width', 'Taper_Head_Width_Meas',
      'Taper_Tail_Width_Meas', 'Slab Weight', 'Slab Lenght', 'Slab Thickness']]
```

Figura 9. *Variables independientes del modelo.*

Estas resultan las más relevantes después de hacer el análisis de correlación para esta fase del proceso.

```
Accuracy: 0.9147009453131902
Matriz de Confusión:
 [[24669  1750]
 [ 2112 16745]]
Reporte de Clasificación:
              precision    recall  f1-score   support

           0       0.92      0.93      0.93     26419
           1       0.91      0.89      0.90     18857

    accuracy                           0.91     45276
   macro avg       0.91      0.91      0.91     45276
weighted avg       0.91      0.91      0.91     45276
```

Figura 10. *Métricas de desempeño del modelo.*

El modelo pudo clasificar bien de acuerdo a la media de tiempo en FM. Se refleja en la precisión, y en el balance de f1-score. Para ambos casos se tienen más de 18,000 evidencias de soporte, por lo que se concluye que el modelo tuvo un buen desempeño.

4.3 Random Forest

Se utilizó un modelo Random Forest para predecir si el tiempo de un planchón en la etapa de finishing mill (FM Thread Time) era mayor que el promedio de este tiempo, lo que puede estar relacionado con un desempeño adecuado o inadecuado del proceso.

```
X = data[['FM Thread Speed', 'FM Final Speed', '1st Accel Time', 'FM Travel Pacing Time',
          'R2 Pacing Time', 'steel_type', 'Slab Width', 'Taper_Head_Width_Meas',
          'Taper_Tail_Width_Meas', 'Slab Weight', 'Slab Lenght', 'Slab Thickness']]
y = (data['FM Thread Time'] > data['FM Thread Time'].mean()).astype(int)
```

Figura 11. *Variables del modelo.*

En este caso, se entrenó un modelo con 100 árboles de decisión para predecir si FM Thread Time sería mayor o menor que el promedio, basándose en las variables predictoras seleccionadas. El conjunto de datos se dividió con 80% de los datos para entrenar el modelo y un 20% para evaluar su desempeño.

```
Accuracy: 0.9529110345436876
Matriz de Confusión:
 [[6417  297]
 [ 236 4369]]
Reporte de Clasificación:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96      6714
           1       0.94      0.95      0.94      4605

    accuracy                           0.95     11319
   macro avg       0.95      0.95      0.95     11319
weighted avg       0.95      0.95      0.95     11319
```

Figura 12. *Métricas de desempeño.*

La precisión del modelo nos indica que fue capaz de predecir correctamente el 95.29% de los casos, además las métricas de evaluación para cada clase indican que el modelo es altamente efectivo para predecir Fm Thread Time en función de las características seleccionadas.
Las métricas de evaluación, como el f1-score y la precisión, fueron consistentes en ambas clases (tiempos dentro del rango esperado y tiempos prolongados).

4.4 KNN

Para este modelo se usaron las mismas variables independientes y dependientes.

```
X = data[['FM Thread Speed', 'FM Final Speed', '1st Accel Time', 'FM Travel Pacing Time',
        'R2 Pacing Time', 'steel_type', 'Slab Width', 'Taper_Head_Width_Meas',
        'Taper_Tail_Width_Meas', 'Slab Weight', 'Slab Lenght', 'Slab Thickness']]
y = (data['FM Thread Time'] > data['FM Thread Time'].mean()).astype(int)
```

Se dividió en conjunto de entrenamiento (80%) y prueba (20%), se escalaron los datos y se probaron 3 K diferentes: 3, 5 y 10. Con base en las métricas de desempeño obtenidas se observa que la mejor k es la K=5, esto lo confirmamos en una gráfica donde observamos cómo cambia la precisión con diferentes valores de k y K=5 tenía la precesión más alta.

```
Resultados para K=5:
[[6332  382]
 [ 293 4312]]
              precision    recall  f1-score   support

           0       0.96      0.94      0.95      6714
           1       0.92      0.94      0.93      4605

    accuracy                           0.94     11319
   macro avg       0.94      0.94      0.94     11319
weighted avg       0.94      0.94      0.94     11319
```

Figura 13. *Matriz de confusión*

Los resultados de este modelo fueron los siguientes:

Para la clase 0 (FM Thread Time por debajo de la media) se obtuvo una precisión de 0.96, esto significa que el 96% de las veces que el modelo predijo esta clase, estaba en lo correcto. Recall de 0.94, el modelo identificó el 94% de los casos donde realmente sucedía esta clase. F1-score de 0.95 , el modelo tiene un buen rendimiento para identificar casos de esta clase.

Para la clase 1 (FM Thread Time por arriba de la media) se obtuvo una precisión de 0.92, esto significa que el 92% de las veces que el modelo prefijo esta clase, estaba en lo correcto. Recall de 0.94, el modelo identificó el 94% de los casos donde realmente sucedía esta clase. F1-score de 0.93, el modelo tiene un buen rendimiento para identificar casos de esta clase.

En la matriz de confusión se observan pocos falsos negativos y falsos positivos en comparación con los verdaderos positivos y los verdaderos negativos.

4.5 ANOVA e Intervalos de confianza

Realizamos dos ANOVAs, la primera fue para la variable 'FM Thread Time' según 'steel_type', aquí obtuvimos un F-statistic alto y un p value menor a 0.05, lo que nos demuestra que hay diferencias significativas de FM Thread Time en los diferentes tipos de acero. Entonces el tipo de acero sí influye en qué tanto tiempo estará el planchón en FM

El segundo ANOVA fue comparando las medias de las predicciones de los modelos: regresión múltiple, regresión logística y random forest. Aquí nuevamente el F-statistic fue alto y tuvimos un p value menor a 0.05. Por lo que concluimos que hay diferencias estadísticamente significativas entre las predicciones de los tres modelos.

Los intervalos de confianza del random forest mostraron que FM Thread Speed y FM Travel Pacing Time son las variables que más influyen en los cuellos de botella porque su intervalo es más estrecho.

Los intervalos de confianza de la regresión logística mostraron que todas las variables de este modelo tienen intervalos de confianza que no incluyen el valor 0, lo que significa que son estadísticamente significativas.

La evaluación en conjuntos de entrenamiento y prueba mostró que no hay un subjuste ni un sobreajuste en los modelos de regresión lineal múltiple y regresión logística.

Conclusiones:

- El modelo con mejor desempeño fue un modelo de clasificación, y fue el random forest con una precisión de 95%, y un f1-score (media armónica de precisión y recall) de 0.96 para la clase 0 y de 0.91 para la clase 1
- Las variables con mayor correlación con FM Thread Time fueron: FM Thread Speed, FM Final Speed y 1st Accel Time.
- FM Thread Speed y FM Travel Pacing Time son las variables que más influyen en los cuellos de botella porque su intervalo es más estrecho, pero el orden de steel_type y los patrones de medidas de los planchones son las variables que podrían controlar.

Conclusiones Individuales

Alejandro: Se detectaron de forma preliminar los impactos que tienen las características inherentes de los planchones en las demoras dentro del FM. Las demás variables, relacionadas a sus velocidades no están en control de Ternium, por lo que considero necesario hacer simulaciones de escenarios cambiando los valores del steel type, o de sus dimensiones para ver cómo se comportan los tiempos dentro del FM. Posterior a ello, se filtra con aquellos que se consideran demoras y con ello podremos llegar a una conclusión sobre este proceso. Esta modelación fue bastante valiosa pues sentó las bases del análisis con modelos de aprendizaje supervisado y los posteriores con no supervisados.

Ania : Este análisis nos dio información sobre los factores que influyen en la etapa FM, los resultados de los modelos y de ANOVA nos ayudaron a resaltar las variables con mayor impacto como la velocidad y la aceleración que es de las que tienen mayor peso en la predicción de los modelos y tienen una correlación alta con FM Thread Time. Con ANOVA se demostró diferencias estadísticamente significativas en FM Thread Time según el tipo de acero, lo que significa que la familia de acero afecta en el tiempo de los planchones durante esta etapa. Dado que la familia del acero tiene un impacto en el tiempo, podría ser útil tener un sistema de clasificación previa que ayude a ajustar los parámetros del proceso en función del tipo de acero, de esta manera minimizar posibles cuellos de botella asociados a esta variable.

Nayeli: Esta evidencia nos permitió en primer lugar definir qué problemática específica queríamos abordar. Limpiamos y armamos nuestra Base de Datos desde 0 pero ya teniendo un enfoque. La matriz de correlación de FM Thread Time nos dio hallazgos de las variables más correlacionadas con el tiempo en que los planchones duraban en esta fase. Los modelos que realizamos nos permitieron predecir y clasificar con una precisión el tiempo de duración en esta fase y si el tiempo sería mayor o menor a la media. Si un planchón tarda menos de la mitad del tiempo promedio en esta Fase se podría inferir que no tuvo un cuello de botella. Si bien las variables significativas de los modelos no eran

variables determinísticas o que estuvieran en total control por parte de Ternium, identificamos que al hacer escenarios de patrones de planchones por familias por medidas específicas podríamos proponer el mejor patrón de ingreso de planchones para disminuir los cuellos de botella en la fase FM. El Anova y los intervalos de confianza nos ayudaron a confirmar las variables significativas de los modelos y el hecho que algunos tienen mejor precisión que otros.

Karla: Considerando que Ternium busca reducir los cuellos de botella en la fase del Finishing Mill decidimos enfocarnos en predecir el tiempo de un planchón en esta fase para las interrupciones. A partir de este trabajo fue posible definir la problemática y generar hipótesis sobre las interrupciones en FM. Los resultados muestran que las características de cada planchón tienen un impacto en el tiempo de la fase, comprobando mediante el Análisis de Varianza, se concluye que el tipo de acero influye en el tiempo de la fase por lo que en próximas entregas será interesante evaluar características dimensionales para la reducción de interrupciones en Finishing Mill teniendo como posibles accionables el orden en el que ingresan los planchones dependiendo de su composición y características. Así mismo, es importante destacar que las variables explicativas para el tiempo de las interrupciones en Finishing Mill corresponden a factores que derivan de fases previas por lo que lo más recomendable será realizar un análisis de cada fase del proceso para identificar aquellas variables sobre las que Ternium pueda influir para reducir las interrupciones.