Análisis de datos sobre producción en la industria textil aplicando la metodología CRISP-DM

Introducción a la Ciencia de Datos y sus Metodologías

Ariel D. López Cota
Fernando Luna Ponce
Elaine Grenot Castellano
Osiris A. Izaguirre Salazar
Rodrigo I. González Valenzuela

1. Comprensión del negocio

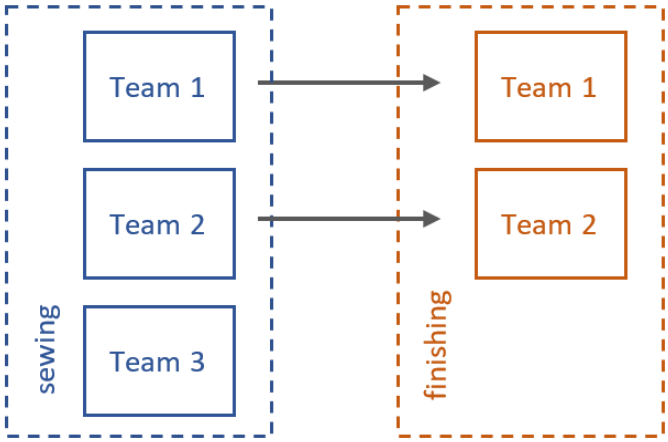
1.1. Objetivos de la línea de investigación

A. Contexto

La producción textil es una de las actividades industriales más demandantes a nivel global, y una pieza clave en la economía bengalí. Sus procesos requieren de muchos procesos manuales y una mano de obra intensiva. [1]

Muchas multinacionales occidentales utilizan mano de obra en Bangladesh, uno de los países más baratos del mundo. Cuatro días son suficientes para que el CEO de una de las cinco marcas textiles más importantes del mundo gane lo que una trabajadora de la confección de Bangladesh ganará durante su vida. [2]

El pipeline de producción que representa nuestros datos consta de algunos procesos, divididos en departamentos, llevados a cabo en actividades secuenciales desempeñadas por distintos equipos. [3]



Para mantener una producción exitosa, la gerencia ha establecido un valor de productividad objetivo (*targeted productivity*) para cada equipo y ha recabado características (incluyendo el porcentaje de productividad real) relacionadas al proceso correspondientes a un periodo aproximado de cinco semanas.

B. Objetivo del negocio

Objetivo Concreto :

Este estudio planea resolver el problema de la ineficiencia en la productividad en la industria de manufactura textil en una fábrica de Bangladesh creando un modelo predictivo utilizando métodos de regresión lineal.

Encontrar las variables relevantes en base a la base de datos proporcionados para la predicción de la productividad para entrenar el modelo y sobre este poder basar los métricos objetivos que se plantea la empresa.

Descripción:

Al desarrollar una línea de producción, el ingeniero industrial debe establecer la productividad objetivo de manera precisa y basada en datos para que la brecha entre la productividad prevista y real sea la mínima.

Si ellos pueden predecir la productividad real de los equipos de trabajo y establecer el objetivo de acuerdo con la predicción, entonces ciertamente pueden minimizar la pérdida financiera, maximizar la ganancia y aumentar la eficiencia de producción.

C. Criterio de éxito (Pendiente)

Crear un modelo que prediga de manera acertada la productividad basado en factores históricos conocidos tales que nos permitan alimentar el modelo con datos a predecir para poder cerrar el gap entre el la productividad objetivo y la productividad real para reducir las ineficiencias y aprovechar de mejor manera los recursos de la compañía.

1.2. Valoración de la situación actual

A. Requisitos, supuestos y restricciones

Se requiere una base de datos con características consideradas como significativas por el ingeniero industrial que se encuentra a cargo de los procesos a eficientar.

Debido a que la variación de la productividad es muy grande teniendo tanto días muy productivos como muy poco productivos se puede presuponer que existen factores culturales que afectan a la productividad.

Se cuenta con una limitante para este estudio, la cual consiste en el tamaño de la muestra de nuestra base de datos teniendo tan solo 1197 observaciones del proceso por lo que si bien es una cifra considerable no promete una precisión demasiado alta.

https://github.com/roglz/produccion_textil_CRISP-DM/blob/main/data/raw_dataset.csv

B. Terminología

- Data Mining (DM): La extracción de información predecible escondida en grandes bases de datos.
- Dataset: Conjunto de datos, pueden ser ordenados o no, pueden ser completos o incompletos.
- Finishing: *Del inglés*, refinamiento. Se refiere a una etapa del proceso de producción textil.
- Sewing: *Del inglés*, de coser o costura. Se refiere a una etapa del proceso de producción textil.

C. Costos y beneficios

Si bien no es posible estimar el Costo ni beneficios directos de la aplicación de este proyecto de ciencia de datos ya que no se planea llegar hasta la etapa de implementación en la industria, es posible darse cuenta que al obtener un modelo predictivo para la productividad total se pueden crear muchos proyectos de ahorros y eficiencias al tomar mejores decisiones al momento de fijar las metas de la industria.

1.3.

A. Objetivo del DM

Encontrar variables con mayor relación a la variable a predecir en este caso la Productividad Real, además de limpiar los datos y organizarlos de tal manera que generen información valiosa para la alimentación del modelo de predicción para de esta manera obtener resultados más apegados a la realidad.

B. Criterios de éxito del DM

El criterio de éxito definido por el equipo para la etapa de DM consiste en la generación de un Dataset en formato “Tidy” o “Gold” el cual nos sirva como entrada para el entrenamiento y validación del modelo predictivo de regresión lineal creado.

1.4. Realizar plan del proyecto

A. Plan del proyecto

- Elaboración de plan de trabajo
- Localización de la fuente informacion
- Descarga y limpieza de los datos
- Análisis exploratorio de los datos
- Detección de variables que influyen en producción
- Análisis de variables y graficación
- Creación de modelo predictivo

B. Evaluación inicial

No es un gran conjunto de datos, a simple vista se detectaron falta de formato de fechas y valores muy altos en comparación al promedio.

2. Comprensión de los datos

2.1. Recolectar los datos iniciales

A. Informe de recolección de datos

Para el proyecto, fueron utilizados datos recopilados por el departamento de Ingeniería Industrial (*Industrial Engineering IE*) de una unidad de fabricación de prendas de vestir de una empresa de renombre ubicada en Bangladesh [3]. No se cuenta con información sobre el proceso de recolección de datos, ni características específicas de la producción adicionales.

Los datos recopilados van desde el 1 de enero de 2015 hasta el 11 de marzo de 2015. El *dataset* tiene una estructura de tabla, con 1197 registros y 15 características.

2.2. Descripción de los datos

A. Informe de descripción de datos

Attribute	Description
date	Date in MM-DD-YYYY
department	Associated department with the instance
team no	Associated team number with the instance
no of workers	Number of workers in each team
no of style change	Number of changes in the style of a particular product
targeted productivity	Targeted productivity set by the authority for each team for each day
smv	Standard Minute Value, it is the allocated time for a task
wip	Work in progress. Includes the number of unfinished items for products
over time	Represents the amount of overtime by each team in minutes
incentive	Represents the amount of financial incentive (in BDT) that enables or motivates a particular course of action
idle time	The amount of time when the production was interrupted due to several reasons
idle men	The number of workers who were idle due to production interruption
actual productivity	The actual productivity value which ranges from 0.0 to 1.0

Revisando la estructura vemos que las fechas no están en el formato más apropiado (fechas), el número de trabajadores no son de tipo entero, cuando debería, y existen datos faltantes para la columna '*wip*'.

```
## Rows: 1,197
## Columns: 15
## $ date                   <chr> "01/01/15", "01/01/15", "01/01/15", "01/01/15", ~
## $ quarter                <chr> "Quarter1", "Quarter1", "Quarter1", "Quarter1", ~
## $ department             <chr> "sweing", "finishing ", "sweing", "sweing", "swe~
## $ day                    <chr> "Thursday", "Thursday", "Thursday", "Thursday", ~
## $ team                   <int> 8, 1, 11, 12, 6, 7, 2, 3, 2, 1, 9, 10, 5, 10, 8,~
## $ targeted_productivity <dbl> 0.80, 0.75, 0.80, 0.80, 0.80, 0.80, 0.75, 0.75, ~
## $ smv                    <dbl> 26.16, 3.94, 11.41, 11.41, 25.90, 25.90, 3.94, 2~
## $ wip                    <int> 1108, NA, 968, 968, 1170, 984, NA, 795, 733, 681~
## $ over_time              <int> 7080, 960, 3660, 3660, 1920, 6720, 960, 6900, 60~
## $ incentive              <int> 98, 0, 50, 50, 50, 38, 0, 45, 34, 45, 44, 45, 50~
## $ idle_time              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ idle_men               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ no_of_style_change     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ no_of_workers          <dbl> 59.0, 8.0, 30.5, 30.5, 56.0, 56.0, 8.0, 57.5, 55~
## $ actual_productivity    <dbl> 0.9407254, 0.8865000, 0.8005705, 0.8005705, 0.80~
```

2.3. Exploración de los datos

A. Informe de exploración de los datos

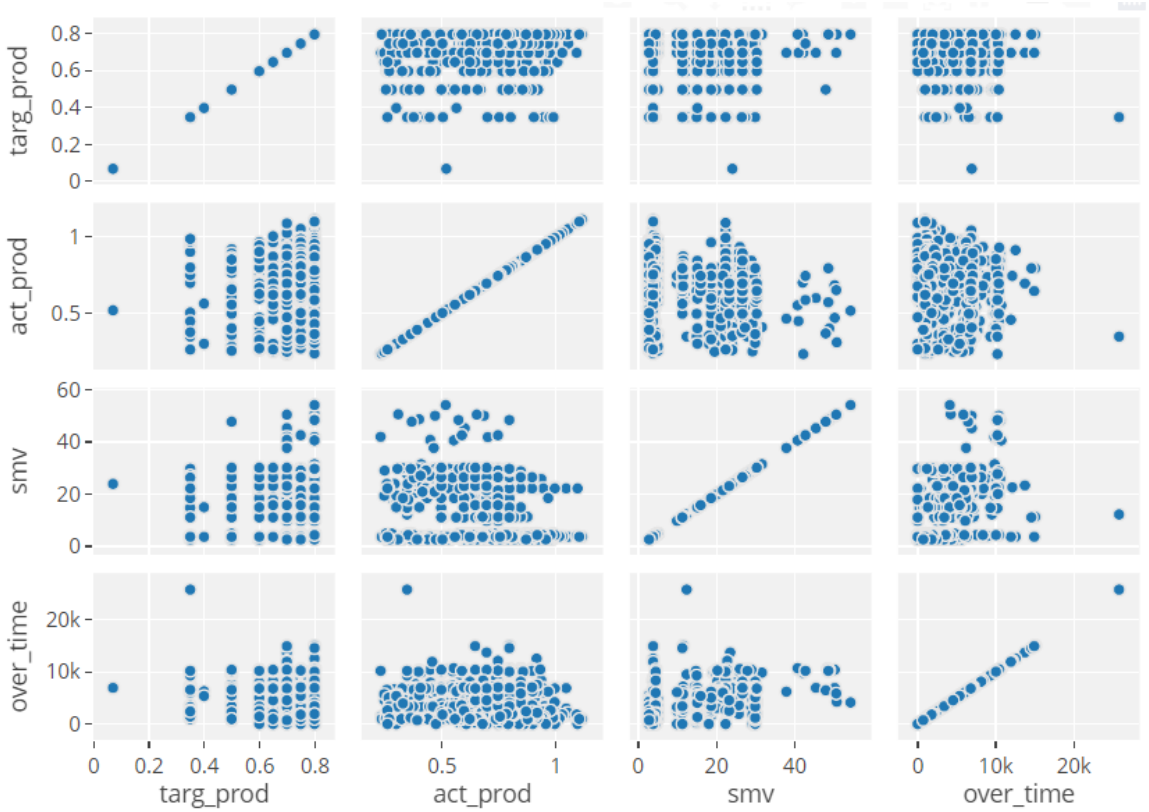
Los datos fueron analizados con el objetivo de verificar su estructura, normalidad y su integridad. Durante el análisis detectamos lo siguiente

- Fechas no en formato correcto
    - La variable *'date'* no contiene un estándar, teniendo formatos dd-MM-yy, mm-dd-yy, dd-mm-yyyy.
- Valores faltantes en variable *'wip'*.

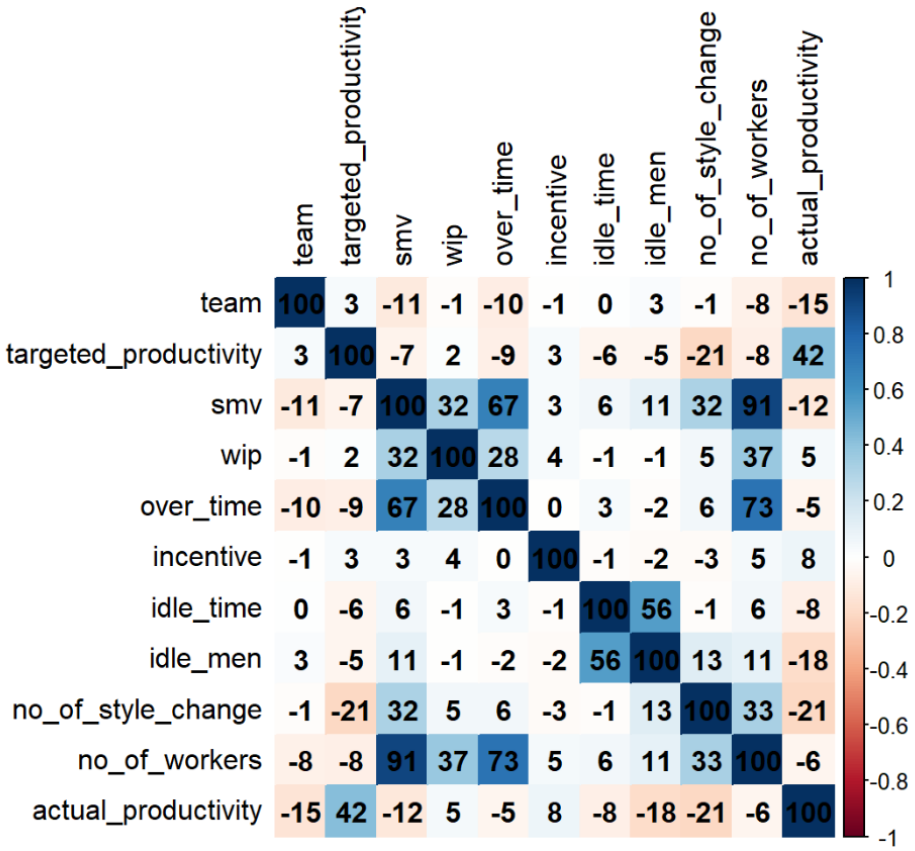
```
data %>%
  sapply(function(no_NA) sum(is.na(no_NA))) %>%
  kbl() %>%
  kable_styling() %>%
  scroll_box(width = "100%", height = "200px")
```

	x
targeted_productivity	0
smv	0
wip	506
over_time	0

a. Visualización de los cada uno de las variables



b. Diagrama de correlación de variables



2.4. Verificar la calidad de los datos

A. Informe de calidad de los datos

Los datos obtenidos aunque no se encuentran en formato “Tidy” si se encuentran suficientemente limpios obteniendo un 2.81% de valores faltantes. Todos ellos en la columna menospreciable *‘wip’*.

3. Preparación de los datos

3.1. Seleccionar los datos

A. Razonamiento inclusión/exclusión

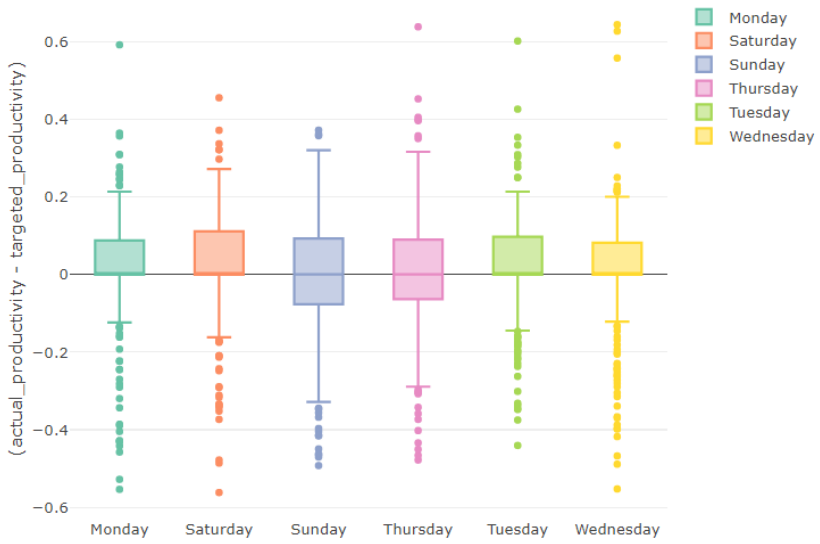
Para la selección de las características a considerar como significativas para entrenar el modelo predictivo se consideró la correlación encontrada entre las diferentes variables cuantitativas con la variable a predecir (en este caso la Producción Real).

- no of Style Change
- Idle Man
- Actual Productivity
- Targeted Productivity

En cuanto a las variables cualitativas se realizó una exploración visual utilizando métodos estadísticos como gráficas de Box así como pruebas de análisis de varianzas (ANOVA):

Producción V Día

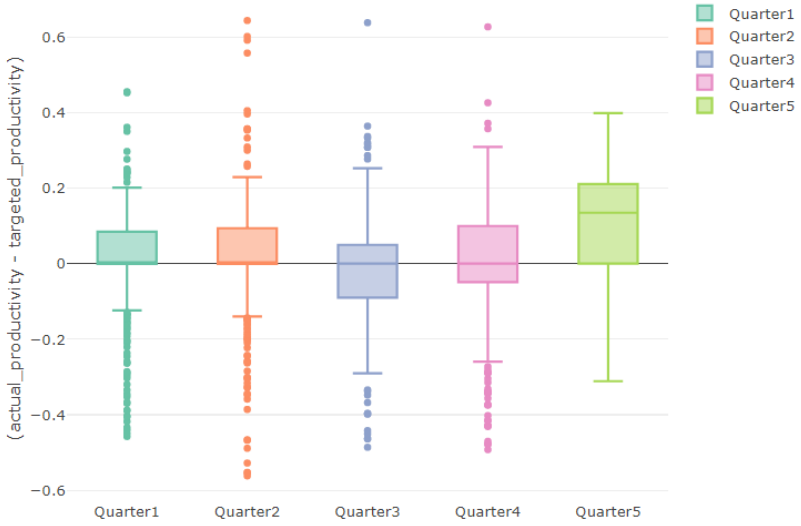
```
d <- data %>%
  plot_ly(
    y = ~(actual_productivity-targeted_productivity),
    color = ~day,
    type = 'box'
  )
d
```



Después de analizar la característica *'day'* podemos ver en el ANOVA que estadísticamente no hay una diferencia significativa pero al analizar la gráfica de Box podemos encontrar ciertos patrones de productividad relacionados a algunos días de la semana en específico por lo que se decide incluir en el modelo.

Producción V Cuarto

```
d <- data %>%
  plot_ly(
    y = ~(actual_productivity-targeted_productivity),
    color = ~quarter,
    type = 'box'
  )
d
```

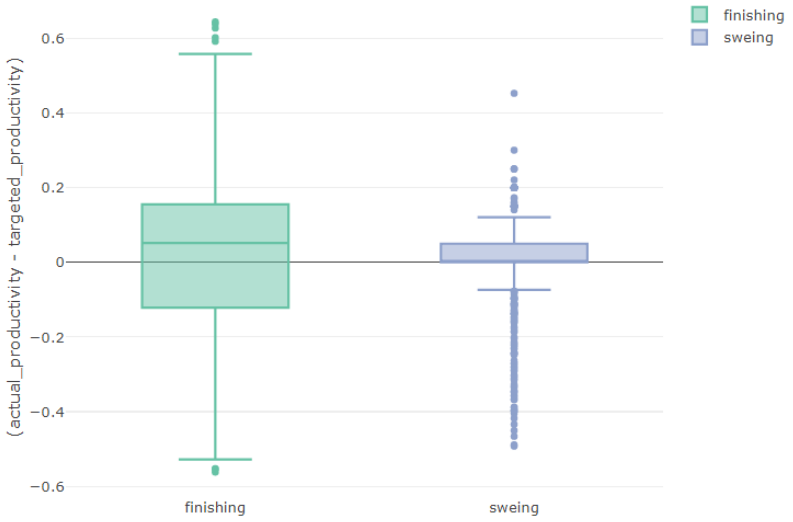


De analizar la característica *'quarter'* se puede observar un patrón de aumento en la productividad los últimos días del mes siendo el Quarter 5 el más productivo de todos con una diferencia en ANOVA muy significativa, por lo tanto se decide implementarlo en el modelo.

Producción V Departamento

```
#Fue necesario modificar los valores de la columna department ya que identificaba dos departamentos finishing dif
erentes debido a que uno se capturó con un espacio al final.
rep_str = c('finishing '='finishing')
data$department <- str_replace_all(data$department, rep_str)

d <- data %>%
  plot_ly(
    y = ~(actual_productivity-targeted_productivity),
    color = ~department,
    type = 'box'
  )
d
```



Por último, la variable cualitativa ‘department’ también nos muestra una diferencia significativa entre sus dos componentes siendo el departamento “finishing” el más productivo, esto se puede entender al conocer la naturaleza de cada uno de los procesos. Por lo anterior, se decide implementarlo en el modelo.

3.2. Limpiar los datos

A. Informe de limpieza de datos

Debido a que los datos se encontraban en su mayor parte limpios desde su captura las únicas modificaciones realizadas para la limpieza de datos fueron las siguientes:

- Modificación de la variable *‘date’* como tipo fecha con formato dd-mm-yy.
- Reemplazo de observaciones NA o NULL con 0 en variable *‘wip’*.
- Reemplazo de observaciones “finishing “ por “finishing” (sin espacio al final) para la estandarización de este departamento en la variable *‘department’*.

3.3. Construir los datos

A. Atributos derivados

En esta sección se realizaron métodos de codificación de variables cualitativas para generar matrices numéricas interpretables por el modelo seleccionado.

Las variables codificadas utilizando el método ‘One_Hot_Encoder’ fueron las siguientes:

- *‘quarter’*:

	quarter_Quarter1	quarter_Quarter2	quarter_Quarter3	quarter_Quarter4	quarter_Quarter5
0	1.0	0.0	0.0	0.0	0.0
1	1.0	0.0	0.0	0.0	0.0
2	1.0	0.0	0.0	0.0	0.0
3	1.0	0.0	0.0	0.0	0.0
4	1.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...
1113	0.0	1.0	0.0	0.0	0.0
1114	0.0	1.0	0.0	0.0	0.0
1115	0.0	1.0	0.0	0.0	0.0
1116	0.0	1.0	0.0	0.0	0.0
1117	0.0	1.0	0.0	0.0	0.0

- *‘department’*:

department_finishing	department_sweing
0.0	1.0
1.0	0.0
0.0	1.0
0.0	1.0
0.0	1.0
...	...
1.0	0.0
1.0	0.0
1.0	0.0
1.0	0.0
1.0	0.0

- *'day'*:

day_Monday	day_Saturday	day_Sunday	...
0.0	0.0	0.0	...
0.0	0.0	0.0	...
0.0	0.0	0.0	...
0.0	0.0	0.0	...
0.0	0.0	0.0	...
...	...	...	...
0.0	0.0	0.0	...
0.0	0.0	0.0	...
0.0	0.0	0.0	...
0.0	0.0	0.0	...
0.0	0.0	0.0	...

- *'team'*:

team_3	team_4	team_5	team_6	team_7	team_8	team_9	team_10	team_11	team_12
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...
0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0

3.4. Integrar los datos

A. Combinación de datos

Para este proceso se realizó la combinación de los dos datasets que se habían manejado por separado utilizando en uno variables cuantitativas y en el otro las cualitativas una vez se realizaron métodos de codificación y escalación.

3.5. Formateo de los datos

A. Datos formateados

One hot code encoding, estandarización.

variables:

1. Quarter
2. Day
3. Department

	quarter_Quarter1	quarter_Quarter2	quarter_Quarter3	quarter_Quarter4	quarter_Quarter5	department_finishing	department_sweing
0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
1	1.0	0.0	0.0	0.0	0.0	1.0	0.0
2	1.0	0.0	0.0	0.0	0.0	0.0	1.0
3	1.0	0.0	0.0	0.0	0.0	0.0	1.0
4	1.0	0.0	0.0	0.0	0.0	0.0	1.0
...	...	...	...	...	...	...	...
1113	0.0	1.0	0.0	0.0	0.0	1.0	0.0
1114	0.0	1.0	0.0	0.0	0.0	1.0	0.0
1115	0.0	1.0	0.0	0.0	0.0	1.0	0.0
1116	0.0	1.0	0.0	0.0	0.0	1.0	0.0
1117	0.0	1.0	0.0	0.0	0.0	1.0	0.0

1118 rows × 31 columns

4. Modelado

4.1 Técnica de modelado

Como se requiere hacer una predicción sobre la productividad se ha seleccionado un modelo de regresión lineal múltiple.

A. Técnica de modelado: Regresión lineal

La regresión lineal es una técnica de modelado estadístico que se emplea para describir una variable de respuesta continua como una función de una o varias variables predictoras. Puede ayudar a comprender y predecir el comportamiento de sistemas complejos o a analizar datos experimentales, financieros y biológicos.[4]

Las técnicas de regresión lineal permiten crear un modelo lineal. Este modelo describe la relación entre una variable dependiente y (también conocida como la respuesta) como una función de una o varias variables independientes

B. Presunciones del modelado

Se quiere realizar la predicción de la productividad en base a la información que tenemos disponible sobre la producción textil.

4.2 Genere plan de prueba

Después de que tenemos nuestros limpios y estandarizados, y una vez que seleccionamos la técnica adecuada para el modelo predictivo, el siguiente paso es la generación del plan de pruebas, para lo cual dividimos nuestros datos, en datos de entrenamiento y datos de prueba en una proporción 80/20

Nuestros datos de entrenamiento los utilizaremos para la construcción de nuestro modelo y los datos de prueba serán los encargados de medir la calidad de nuestro modelo.

4.3 Construir el modelo

Se utilizó la librería sklearn para la generación del modelo de regresión lineal.

```
#Definimos el algoritmo a utilizar
from sklearn import linear_model
lr_multiple = linear_model.LinearRegression()
```

Después de seleccionar el modelo, el siguiente paso fue realizar el entrenamiento del modelo con los datos de entrenamiento.

```
#Entrenamos el modelo
lr_multiple.fit(X_train, y_train)
```

4.4 Evaluar el modelo

Una vez que se entrenó el modelo el siguiente paso es la obtención de parámetros para medir la precisión de este. Se tomaron en cuenta las métricas para los datos de entrenamiento y para los datos de prueba.

	RMSE	MAE	R2
Training	0.1402	0.1020	0.2786
Test	0.1425	0.0996	0.2057

Como se puede apreciar en la tabla, los valores del RMSE y del MAE nos dan una idea de que diferencia existe entre la predicción y el valor real, es decir el error, en este caso vemos que existe un error de cercano al 10%. El R2 nos dice que tan bien se pueden predecir los datos, mientras más cercano el valor a 1, mejor es el modelo, en este caso nos salió un resultado bastante pequeño, por lo que se deben de seguir ajustando los parámetros para obtener un mejor modelo.

5. Evaluación

5.1 Evaluar resultados

A. Valoración de resultados

- Los últimos días del mes hay mayor índice producción
- El departamento de “finishing” tiene mayor proporción de producción objetivo no lograda.

B. Valoración de resultados

- El modelo generado tomando como base las variables que tienen mayor índice de correlación con la variable “targeted productivity”, no presenta un buen grado de exactitud, por lo que sería necesario seguir haciendo pruebas para mejorar el modelo.

5.2 Revisar proceso

A. Revisión del proceso

- Se siguieron correctamente los pasos de la metodología CRISP-DM desde la comprensión del negocio para decidir que lo que se requería hacer, hasta la creación del modelo predictivo. Las tareas en general quedaron bien definidas.

5.3 Determinar próximos pasos

A. Lista de posibles acciones

- Utilizar todas las variables y realizar un análisis de componentes principales para obtener nuevas variables que nos permitan explicar la varianza de nuestros datos utilizando toda la información disponible y reduciendo el número de variables necesarias para la elaboración del modelo predictivo.
- Realizar modelos predictivos individuales para cada departamento existente, “finishing” y “sweing”, con esto se esperaría obtener una predicción más precisa para cada unos de estos departamentos.

6. Implantación

En base a los resultados obtenidos se determinó que no es posible realizar una implementación del modelo como tal ya que el modelo no proporciona suficiente precisión como para que pueda ser implementado en la industria de la confección, se requieren recabar más información sobre los diferentes procesos para obtener un buen modelo predictivo y este pueda implementarse de manera satisfactoria.

Anexos

[1]https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees#

[2]https://es.wikipedia.org/wiki/Industria_textil#Banglad%C3%A9s

[3] Artículo *Deep Neural Network Approach for Predicting the Productivity of Garment Employees*

[4] https://la.mathworks.com/discovery/linear-regression.html