LABORATORIO 02

TEMAS DE INGENIERÍA MECATRÓNICA H

Nombre: Armando Arturo Custodio Díaz

Código: 20196355

Generación de Conclusiones

1. Identificador único

Establecer el identificador único de la data.

```
# PHME_data_train.groupby('PanelID').nunique().reset_index().shape[0]

if PHME_data_train.groupby(['PanelID','FigureID','ComponentID','PinNumber']).nunique().reset_index().shape[0] == PHME_data_train.shape[0]:
  print('identificador unico encontrado!')
else:
  print('no se encontro')

# Esto se ha obtenido dada la información en la pagina:
#   "In this scenario, the SPI data, reports, for each PCB, characteristics such as Identifiers: PanelID, FigureID, ComponentID, PinNumber"
```

Al establecer un identificador único para los datos en el conjunto de datos de Inspección de Pasta de Soldadura (SPI) se compone de cuatro elementos clave: PanelID, FigureID, ComponentID y PinNumber. Estos identificadores forman un conjunto integral que permite una clasificación única y específica para cada PCB en el proceso de fabricación.

2. Gráficos de variables numéricas continuas

Se observa que, al aplicar la regla del 99.7% para eliminar valores atípicos y reemplazarlos con la media, se genera una acumulación en ciertos puntos en los gráficos de densidad para Height%, Volume%, y Area%, para ser más específicos, en los datos de la media aritmética.

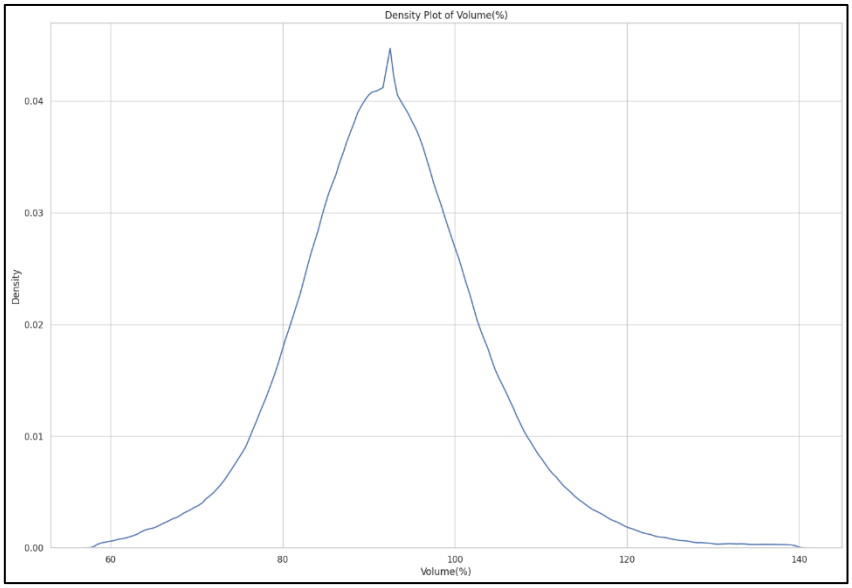


Gráfico de densidad: Volume(%)

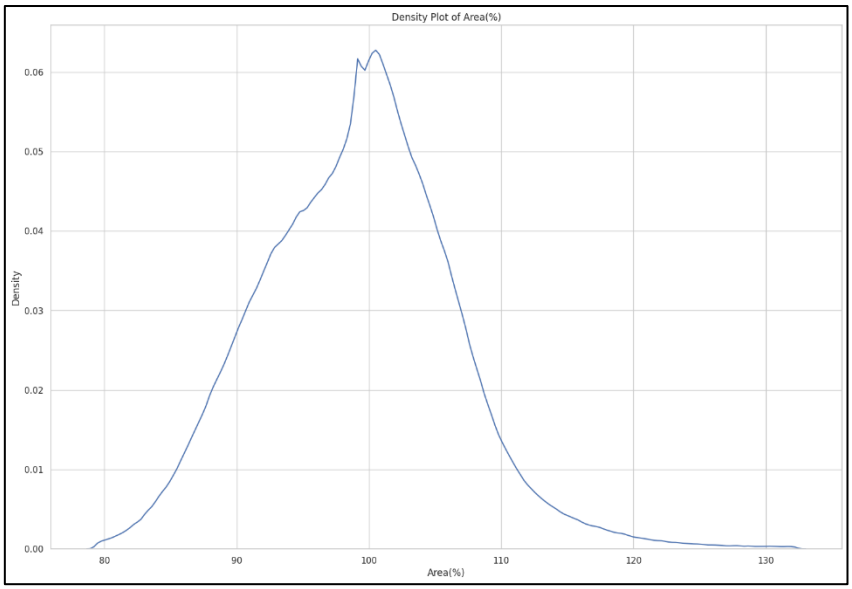


Gráfico de densidad: Area(%)

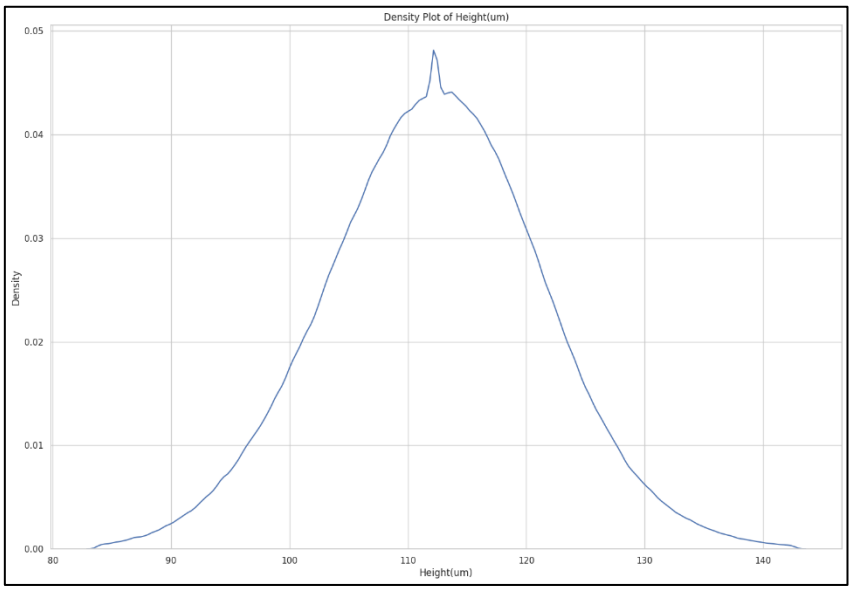


Gráfico de densidad: Height(um)

Esta acumulación se debe a que, al eliminar los valores extremos y acercarlos a la media, se crea una concentración de datos alrededor de ese valor. Este fenómeno es común al aplicar técnicas de manejo de outliers y resalta la importancia de considerar alternativas, como la sustitución por la mediana o transformaciones, para evitar distorsiones en la distribución de los datos.

3. Gráficos de variables categóricas nominales

La distribución observada en el gráfico de frecuencia relativa para PinNumber revela patrones significativos. Los valores 1 y 2 destacan con aproximadamente un 33%, posiblemente indicando una frecuencia común para los primeros pines en el conjunto de datos. Los valores 3, 4, 5 y 6

muestran una frecuencia más baja, alrededor del 5%, sugiriendo que estos pines son menos comunes o específicos para ciertos casos. Sin embargo, llama la atención que los valores desde 33 hasta 48, junto con THERMAL1, tienen una frecuencia considerablemente menor al 1%.

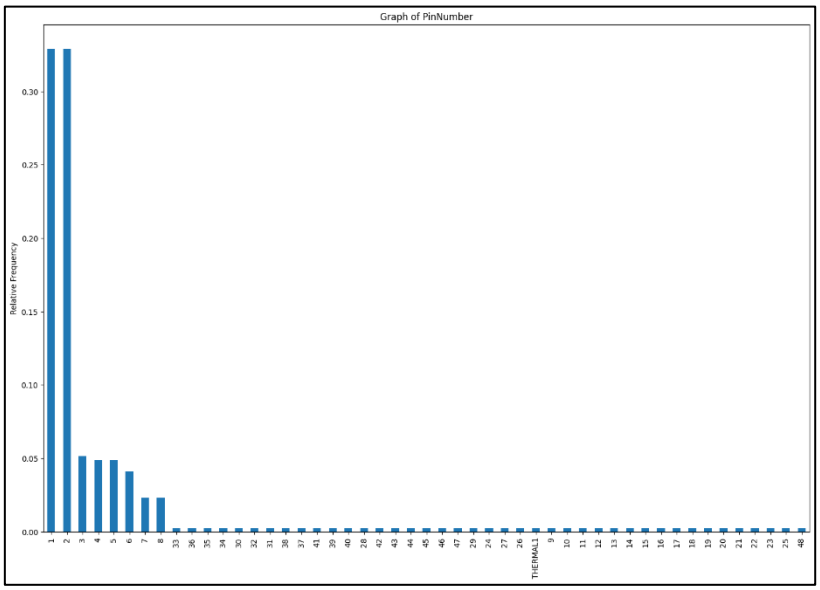


Gráfico de barras de frecuencia relativa: PinNumber

La predominancia abrumadora del valor "GOOD" en más del 95% en la gráfica de frecuencias de la variable RESULT indica que la gran mayoría de los componentes inspeccionados no presentan defectos según la máquina de inspección. Este resultado es positivo y refleja la eficacia del proceso de fabricación. La presencia de un valor ligeramente significativo, como "W.Insuff" después de "GOOD", sugiere que existe una proporción pequeña pero no despreciable de componentes con insuficiencias de soldadura. Las demás denominaciones se pueden despreciar.

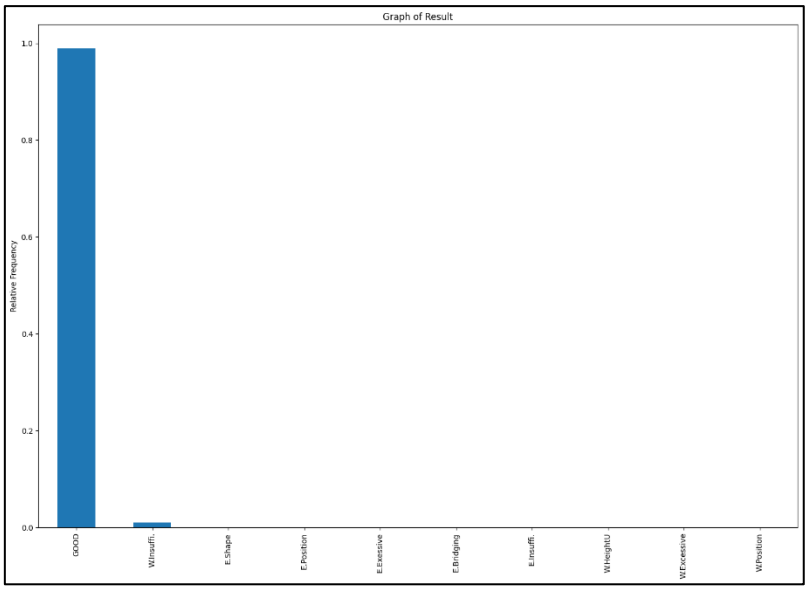


Gráfico de barras de frecuencia relativa: Result

La distribución equitativa del FigureID del 1 al 8 en el conjunto de datos SPI sugiere que cada panel de producción contiene ocho PCBs individuales. Esta estructura organizativa facilita un etiquetado único y específico para cada componente en el proceso de fabricación de placas de circuito impreso.

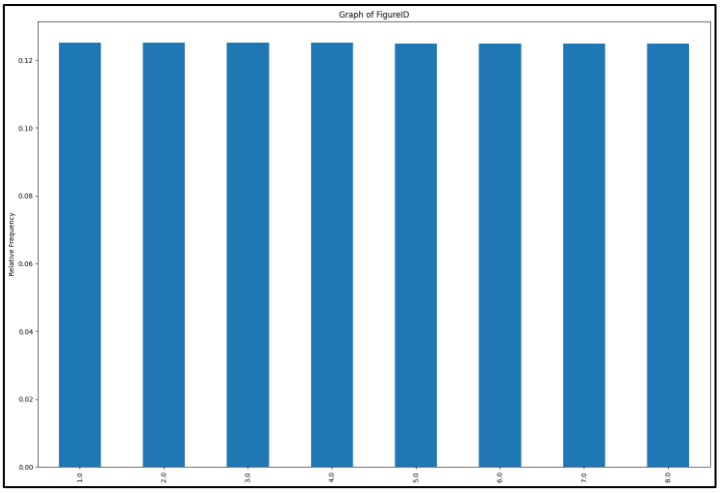
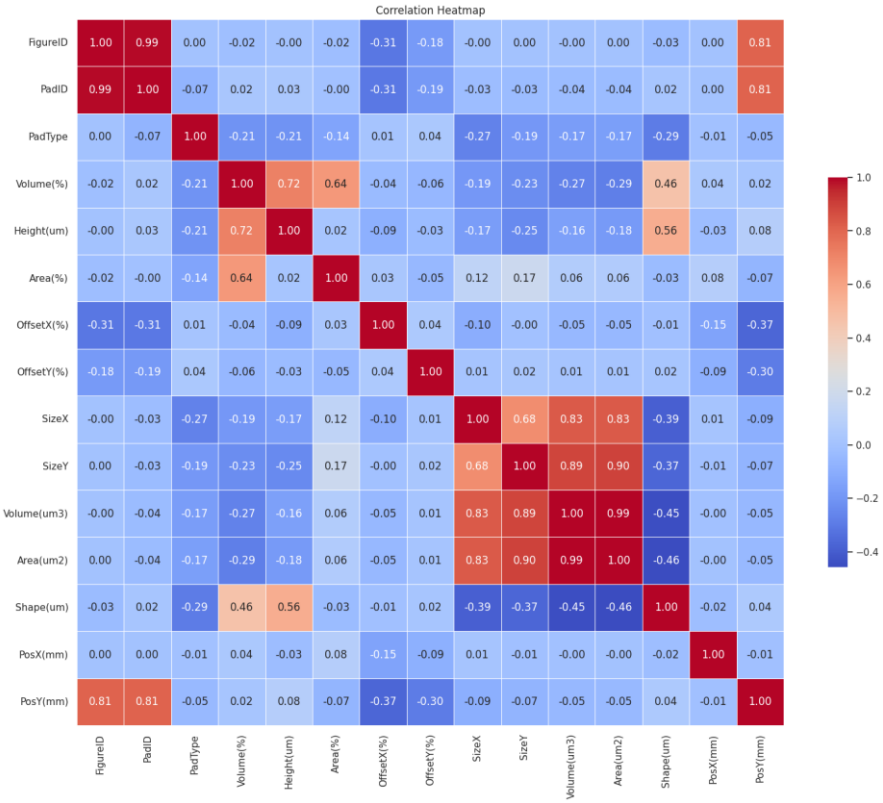


Gráfico de barras de frecuencia relativa: FigureID

Dado el Correlation Heatmap. Se analizarán los casos de alta correlación entre variables, conviene considerar obviar ciertas variables para evitar redundancia cuando se busca simplificar el modelo y mejorar la interpretación de los resultados. La redundancia se presenta cuando dos o más variables están altamente correlacionadas, lo que significa que proporcionan información similar o casi idéntica al modelo.



Correlation Heatmap

- Cuando se analiza la alta correlación de 0.99 entre PadID y FigureID en el conjunto de datos SPI, conviene considerar eliminar uno de estos identificadores para evitar redundancia, ya que ambos están altamente asociados. En este caso, podría ser beneficioso conservar FigureID, ya que podría ser más específico para identificar individualmente cada PCB en lugar del panel general, simplificando el modelo sin perder información crucial.
- En la correlación de 0.83 entre Volumen(um3) y SizeX, así como entre Area(um2) y SizeX es conveniente conservar SizeX, ya que representa una dimensión específica de las características de la pasta de soldadura y podría ser más fácil de interpretar y utilizar en modelos predictivos.
- La correlación perfecta de 0.99 entre Area(um2) y Volumen(um3) indica una redundancia práctica entre estas dos medidas. Se puede eliminar ambas, pues tienen una estrecha relación con SizeX y dicha variable contiene información de Area y Volumen.
- La la correlación de 0.81 entre PosY(mm) y PadID, podría ser beneficioso obviar PadID, lo cual ya se propuso anteriormente

El análisis de correlación en el conjunto de datos SPI es esencial para optimizar futuras aplicaciones de machine learning. Identificar y eliminar variables redundantes mejora la eficiencia y precisión de los modelos al reducir la complejidad y evitar la introducción de información innecesaria. El análisis de correlación es crucial para construir modelos más efectivos y eficientes en el procesamiento de datos y aplicaciones de machine learning subsiguientes.