Box Strength Predictive Model

Team members:

Mike Daly, Supanart Sukrason, Xindi Lu, Rahul Veerapur, Sai Sanjna Chintakunta

Introduction

Many of the items that are purchased by consumers come packaged in boxes of various sizes. If designed correctly, good packaging will promote brand recognition and protect the product from potential damage during shipment and while it is stored on a retailer's shelves. The latter part of this example emphasizes the importance of strength testing because if the box that a product is packaged in structurally fails, it could lead to profit loss and limit the product's availability until issues with the manufacturing of the box are resolved.

This report will investigate the corrugated box manufacturing process and will apply data mining techniques on data provided by SCGP, a packaging manufacturer based in Southeast Asia, to create a model to predict box strength. Analysis of data, such as the strength of raw materials (ECT), feed and printing cylinders gap, and machine speed, will help determine if the box for a particular product will pass the strength test. Based on these parameters in the box manufacturing process, the following objectives have been determined:

- Predict box strength based on the strength value of raw materials (ECT) and the amount of width applied to raw materials going through feeding and printing cylinders, reducing machine downtime and mitigating opportunity losses from the box testing strength process.
- Finding optimal feed gap size and machine speed based on product identifier (SKU).

Background

Machine downtime is a significant concern for the corrugated box manufacturing industry, leading to an annual opportunity loss exceeding $2.7 million. One of the primary contributors to this downtime is the strength testing process for corrugated boxes. During this process, production must halt to await results from the Quality Assurance Laboratory. If the results meet the strength specifications for each SKU, production can resume seamlessly. However, if the results do not meet the required standards, machine operators must rely on their experience to adjust parameters and achieve the desired quality, further delaying production.

The implications of machine downtime extend beyond mere financial loss. Downtime can disrupt supply chains, affect customer satisfaction, and tarnish the reputation of the manufacturing company. In an industry where timely delivery and product quality are paramount, minimizing

downtime is essential for maintaining competitive advantage. Moreover, eliminating inefficiencies in the production process can lead to cost savings, better resource utilization, and increased profitability.

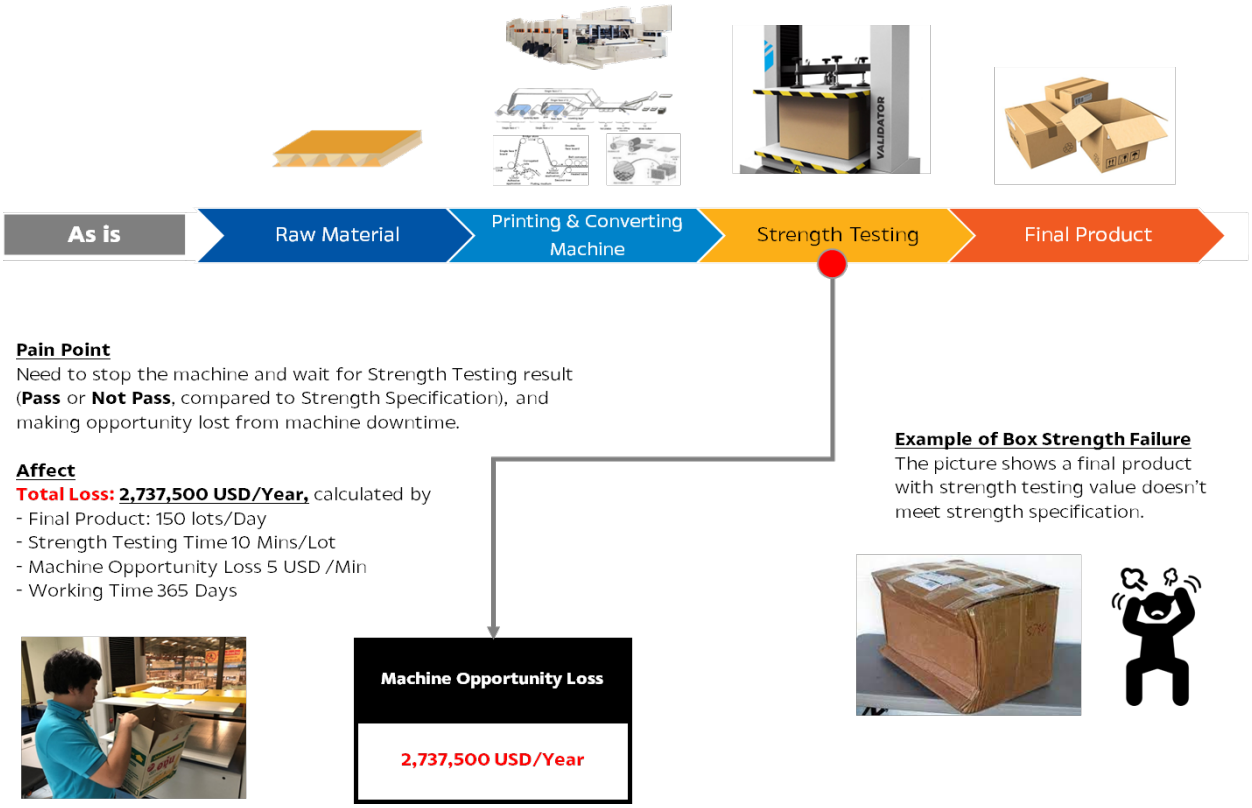


Figure 1: Corrugated Production Process

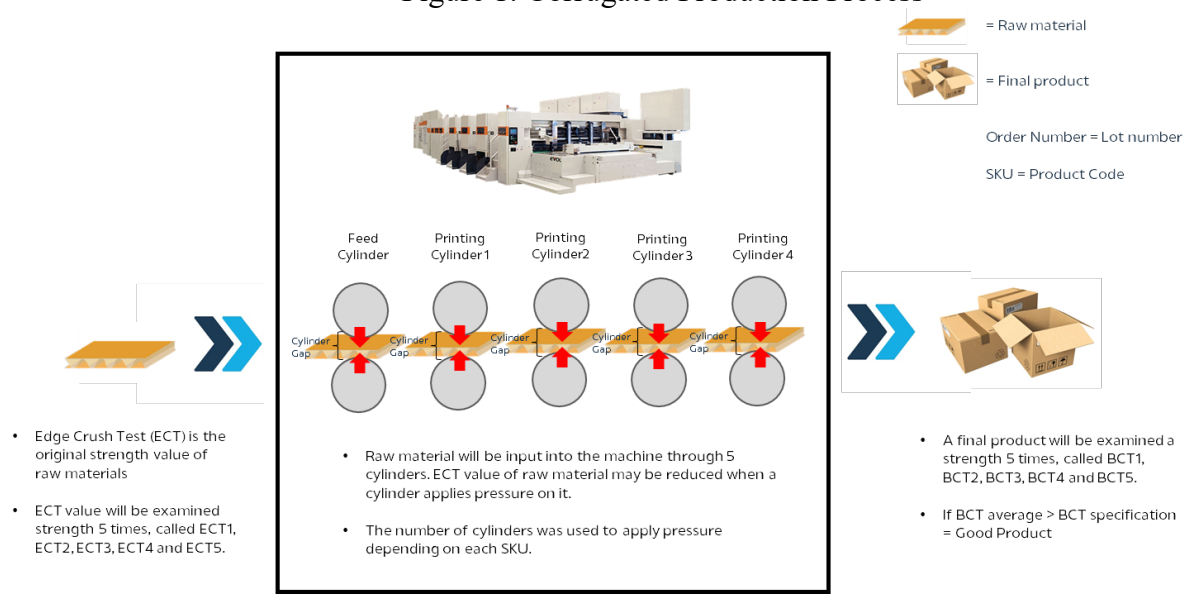


Figure 2: Corrugated Production Machine

Related Data

Related parameters that may affect the strength of the final product (BCT) include the strength of the raw material (ECT), and machine parameters (feed gap, printing cylinder gap, and machine speed). The input parameters that we need to observe are as follows and the picture below:

- Strength of the raw material (ECT): Test to determine the amount of weight that can be stacked on a box without damage.
- Feed Gap: Width applied by a feeding cylinder roll on a raw material during the production process, forming a final product.
- Printing Gap: Width applied by a printing cylinder roll on a raw material during the production process, forming a final product.
- Machine Speed: The rate at which the final product is manufactured.
- Specification of the final product (BCT Spec): Controlled parameter of the strength value of the final product. The actual BCT value from a final product does usually not exceed the BCT Spec.

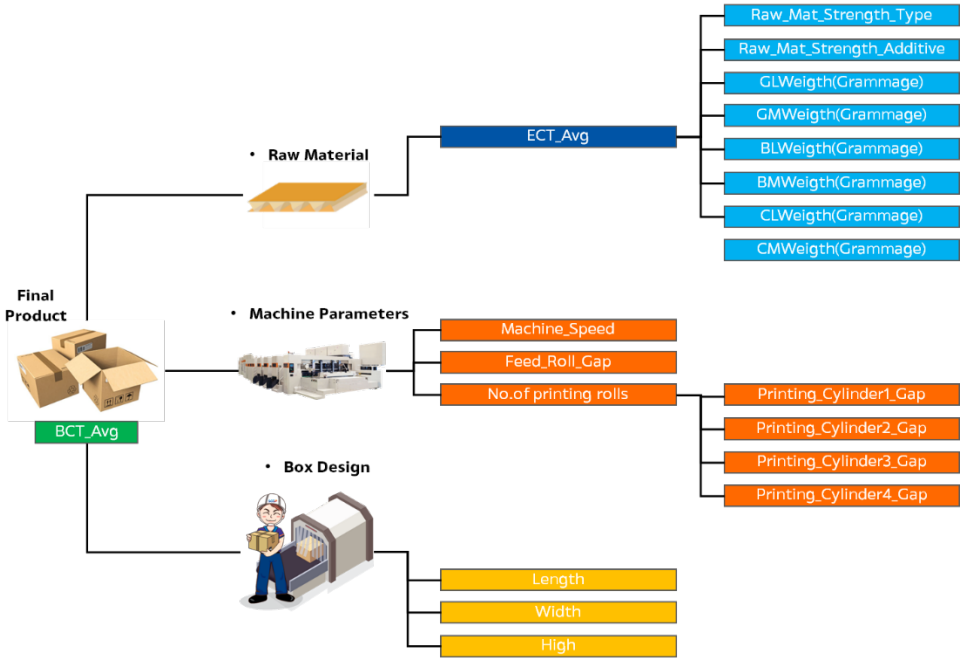


Figure 3: Variable Diagram

Raw Data Preparation

This project leverages two primary data sources to deliver comprehensive insights:

- Machine Parameters: These parameters are meticulously collected from the machine controller systems. They include a variety of operational metrics such as feed gap, printing cylinder gap, and machine speed.
- Raw Material Strength (ECT): ECT is sourced from the supplier's quality laboratory database.

Our project does not use data from the internet or open-source sources. Therefore, we need to cooperate with the engineering teams in Thailand and Japan to install machine data logging and develop automatic data-collecting software called "Speed Link" using C#. This software will gather machine parameter data from the machines during the production process.

Moreover, this software can transform and handle unsuitable data formats from the machine data logging to be stored in the SQL database, reducing the need for some data cleaning. We installed 7 machine data logging units, divided into 2 machine groups include Group 1 (4 Mitsubishi EVOLs) and Group 2 (3 Isowa FALCONs).

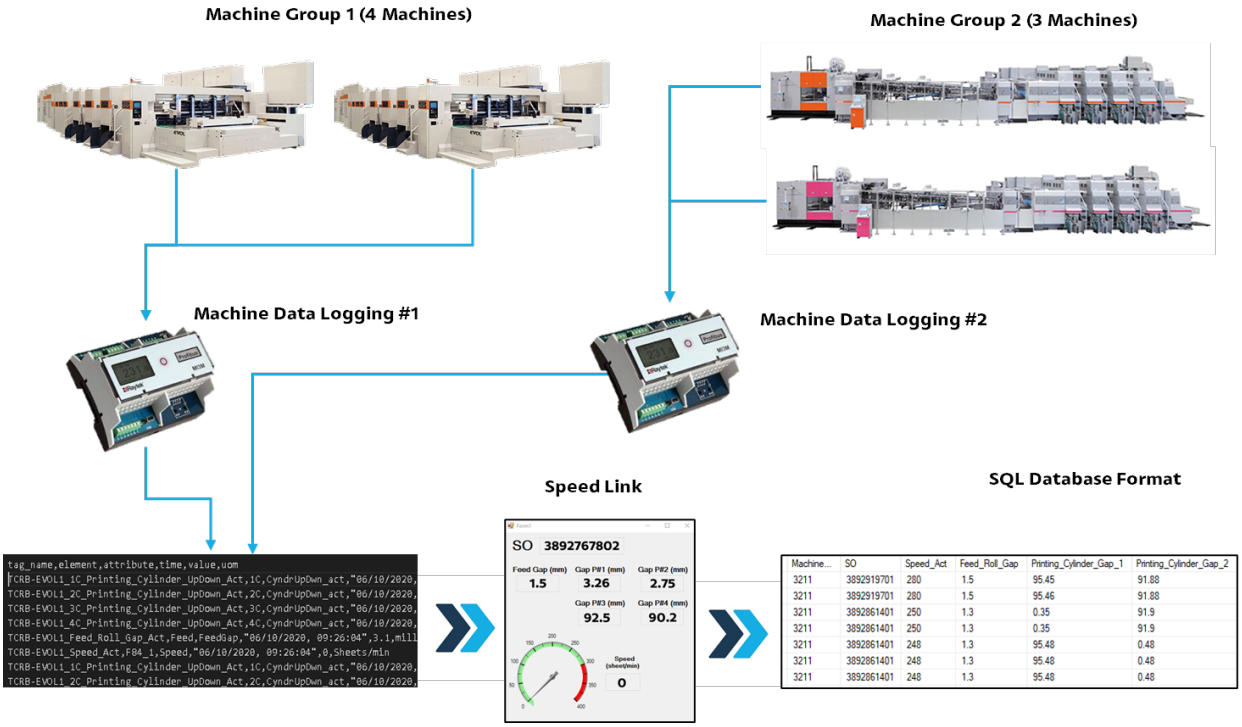


Figure 4: Data Logging Installation and Machine Parameter Data Collecting Diagram

Secondly, we develop an API called "ECT Link" to connect to the supplier quality laboratory database and retrieve raw material strength data. Finally, we merge both data sources

and other product specification data together such as product dimensions, raw material strength additive, and etc. to create a repository that is easily accessible for further analysis.

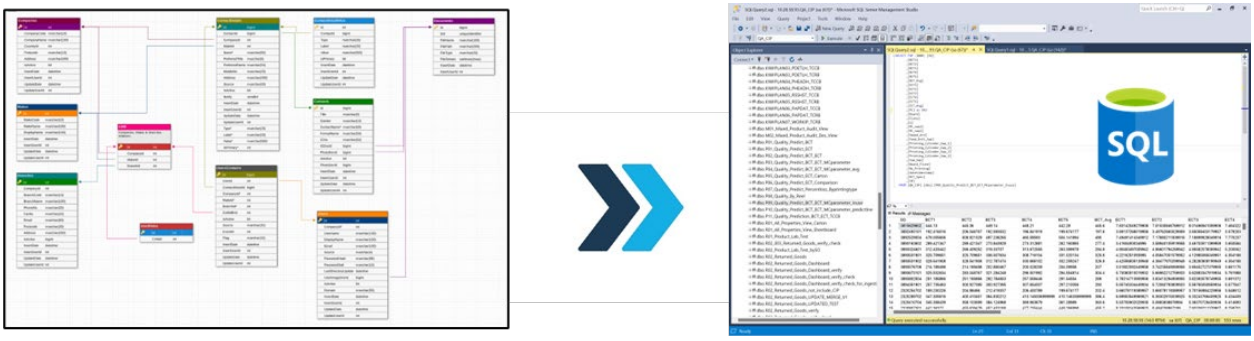


Figure 5: Data Mapping Diagram

Data Cleaning

1. **Dataset Integration:** Merging 2 datasets from 2 different machine groups, Group 1 (4 Mitsubishi EVOL) and Group 2 (3 Isowa FALCON), into one dataset.

2. **Null Cleaning:** Some rows show as Null due to network failure, resulting in incomplete data transfer from the machine controller or the supplier database.

Index	Rwf	Order_Numb	BCT1	BCT2	BCT3	BCT4	BCT5
334	258	3913157101	273.8739	262.071	268.1293	263.9866	261.6953
417	258	3909976305	265.805	266.7601	264.9592	270.2103	262.0092
518		3913738201					
1087		3918280402					
1282		3913697502					
1632		3912296901					
1880	258	3919427202	264.4114	272.0429	271.1536	271.6856	263.3156

Figure 6: Null Data

Applying Null filter to drop off some rows using the example code

- condition = appended_df['Order_Number'].notnull() & appended_df['BCT1'].isnull()
- appended_df = appended_df[condition]

3. **Data Format Transformation:** Some columns need to be transformed from String to Float, making it easy to calculate in the future.

Raw_Mat_Strength_Additive	Raw_Mat_Strength_Additive_Percent
G3	15%
G1	5%
G3	15%
G3	15%
G3	15%
G3	15%
G1	5%

Figure 7: Data Conversion

Applying data transformation condition using the example code

- strength_mapping = {
  'G0': '0.00',
  'G1': '0.05',
  'G2': '0.10',
  'G3': '0.15%',
  'G4': '0.20',
  'G5': '0.25',
  'G6': '0.30' }

- appended_df['Raw_Mat_Strength_Additive_Percent'] = appended_df['Raw_Mat_Strength_Additive'].map(strength_mapping)

4. **Create Binary Column by Assigned Condition:** Generating 4 new binary columns from product reject condition and production standard.

   a. _Result: BCT_Avg < BCT_Spec or %Diff between BCT_Avg and BCT_Spec < 3% = 0 else 1

   b. Finished_Product_Hig : Box with high between 300-600 = 1 else = 0

   c. Machine_Speed: < 350

   d. ECT_Avg: > 1.98

Finished_Product_Hig	Machine_Speed	ECT_Avg	Rejection_Result
0	1	1	1
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0
0	1	1	0

Figure 8: Binary Column

5. **Create Binary Column by Dummy Technique:** Creating 30 new binary columns from the dummy method for categorized columns.

Raw_Mat_Strength_Type_B	Raw_Mat_Strength_Type_BC	Raw_Mat_Strength_Type_C	Raw_Mat_Strength_Additive_G0
1	0	0	1
0	0	1	1
1	0	0	0
0	0	1	0
0	0	1	0
0	0	1	0
1	0	0	1
0	1	0	0
0	0	1	0

Figure 9: Binary Column

6. **Null Value Verification:** Final checking Null value in the dataset.

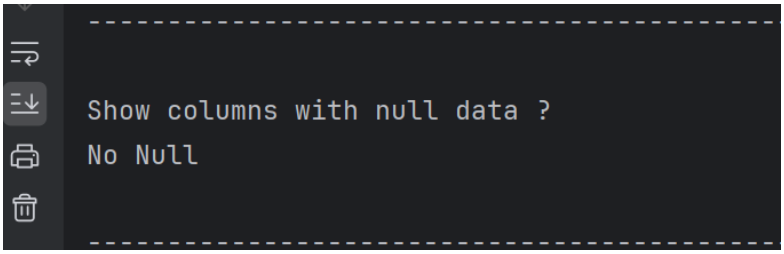


Figure 10: Null Value Result

7. **Null Value Verification:** Generate some descriptive statistics to understand the dataset.

	BCT_Avg	ECT_Avg	Feed_Roll_Gap	New_Printing_Gap_1	New_Printing_Gap_2	New_Printing_Gap_3
count	44641	44641	44641	44641	44641	44641
unique	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	NaN	NaN	NaN
mean	299.43	5.21	1.79	0.47	1.56	2.08
std	113.04	1.59	0.71	0.28	0.85	0.92
min	112.10	2.39	0.09	0.00	0.04	0.10
25%	218.92	4.04	1.34	0.23	0.95	1.40
50%	269.17	4.82	1.75	0.46	1.35	1.91
75%	365.15	6.20	2.18	0.71	2.00	2.56
max	786.65	13.83	4.40	1.00	5.55	5.75

Figure 11: Descriptive Statistics for Each Column

8. **Using Data Visualization:** Data visualization of some major variables, following the ISO standard of corrugated carton production mentioned in the 'Related Data' section, directly affects the strength of the corrugated box. It is evident that machine speed rarely has a relationship with BCT_Avg. We will explore other variables that may have a hidden effect in the next part.

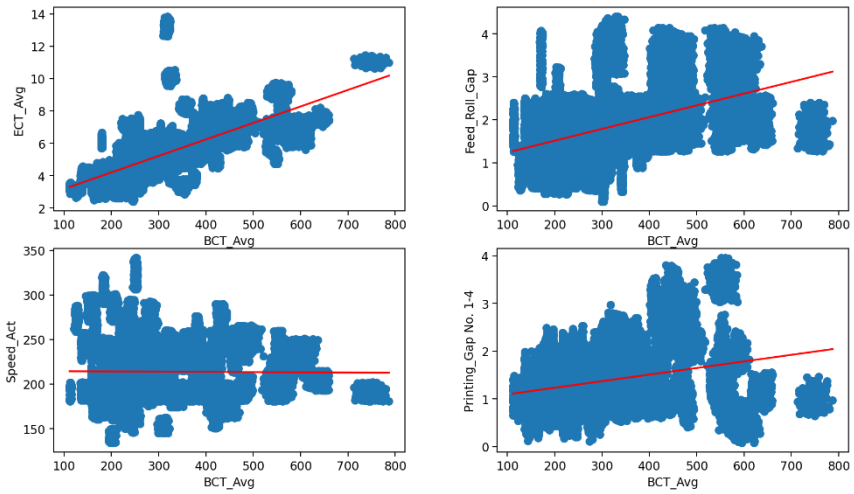


Figure 12: Major Variables Vs Box Strength (BCT_Avg)

In addition, a histogram can help us clearly understand that our dataset is mostly skewed, and some variables, such as Speed_Act, exhibit a bimodal distribution.

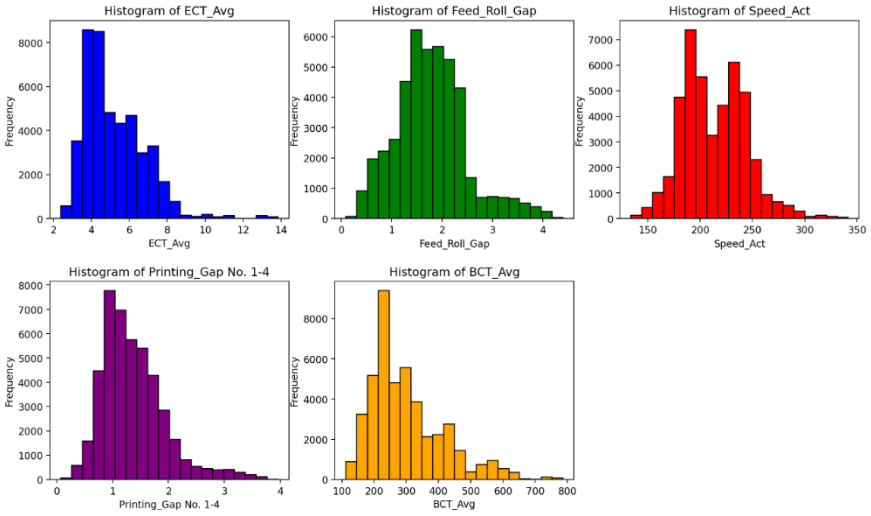
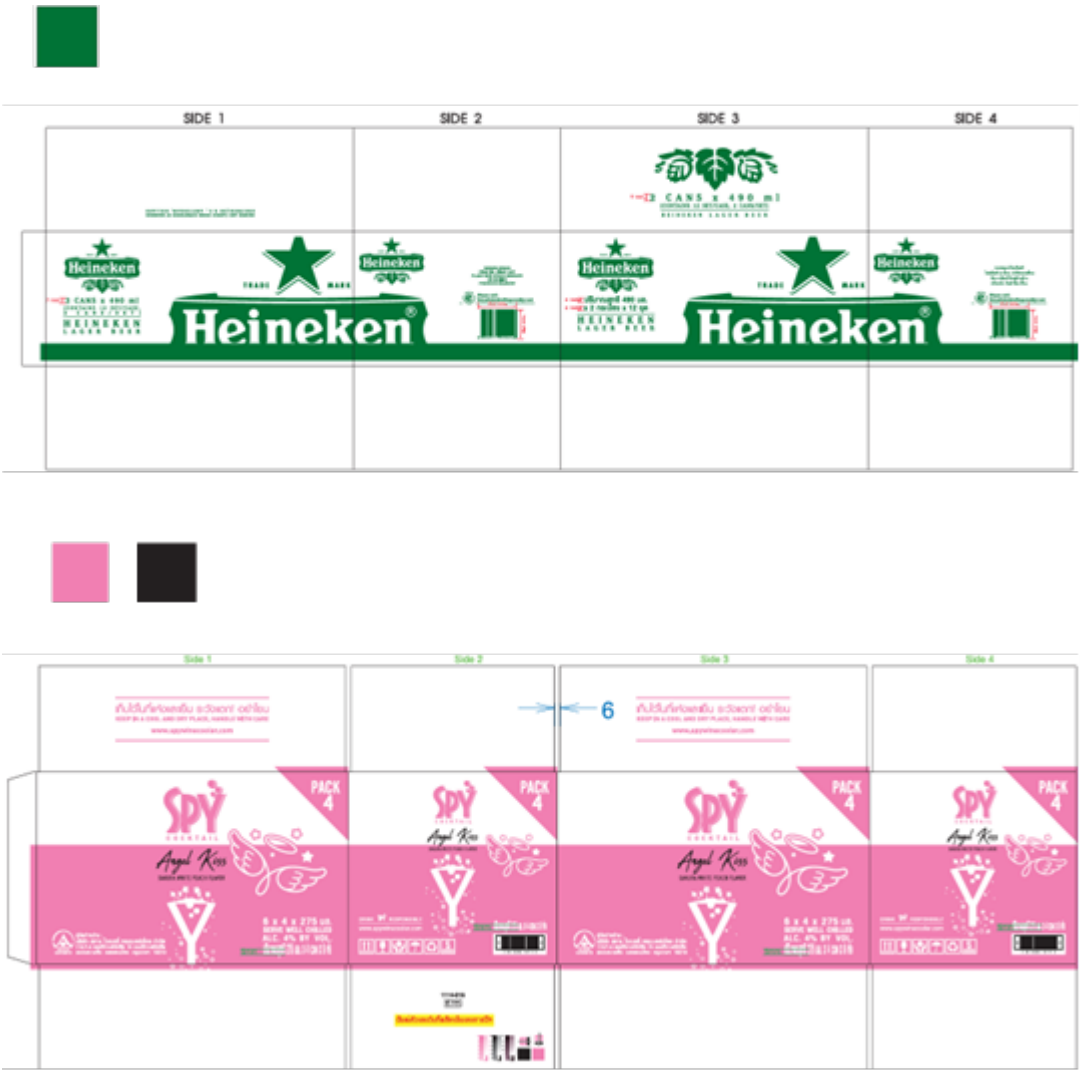


Figure 13: Histogram of Major Variables

9. **Data Cleaning Conclusion:** After we cleaned our data and get the new dataset, the data record was reduced from 125,288 to 124,857.

10. **Dataset Separation:** We divided the data into 5 datasets due to the corrugated production process, which involves the number of printing cylinders being applied to the raw material depending on the printing design. This division allows us to analyze the effect of box strength more deeply.



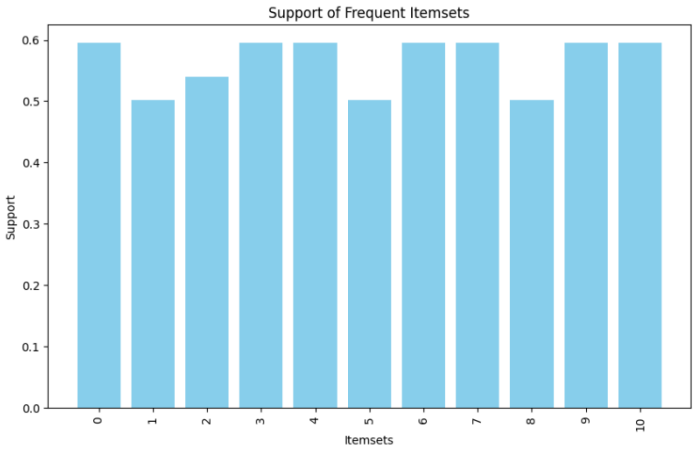
Green: Using 1 Printing Cylinder

Pink + Black: Using 2Printing Cylinders

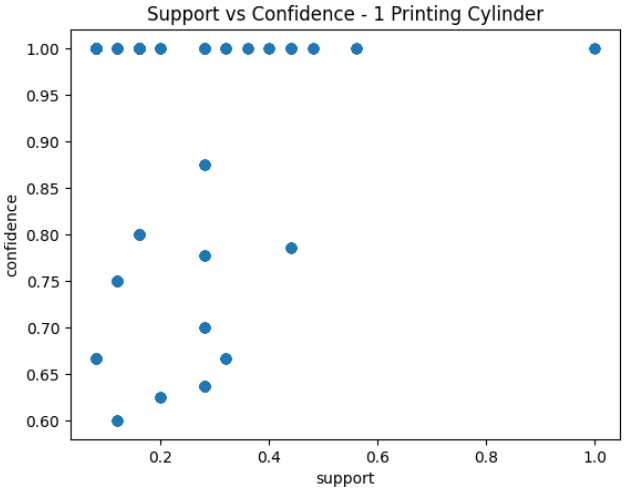
Frequent Item Mining

Apply Apriori and FP-Growth algorithms to the binary columns. Use Rejection_Result=1 and identify significant vaariables impacting the strength of boxes produced. By Varying the minimum support values between 0.3 to 0.5, we aimed to filter the most frequent and relevant patterns from binary columns representing production variables.

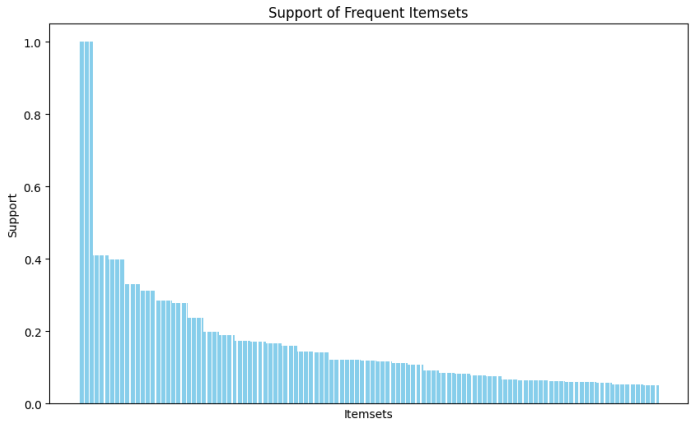
Support vs. Confidence



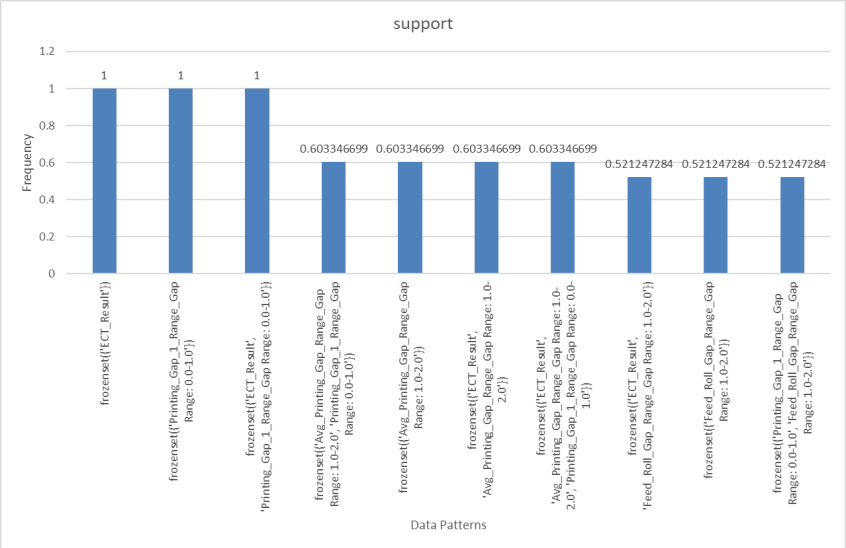
This is the results with 0 printing cylinder. The support values range between 0.5 to 0.6. Some itemsets have the same support values and they are the highest near 0.6. These itemsets occur the most frequently and might represent crucial combinations of conditions that consistently affect the production outcome. There are three itemsets have slightly lower support values, indicating they cooru less frequently than the top itemsets.



The results for 1 printing cylinder have a significant number of rules achieve confidence =1.0, This indicates that these patterns are almost always correct in predicting the outcome. The high-confidence patterns indicate strong associations and controlling the gap in 1 printing cylinder could significantly affect product strength outcomes.



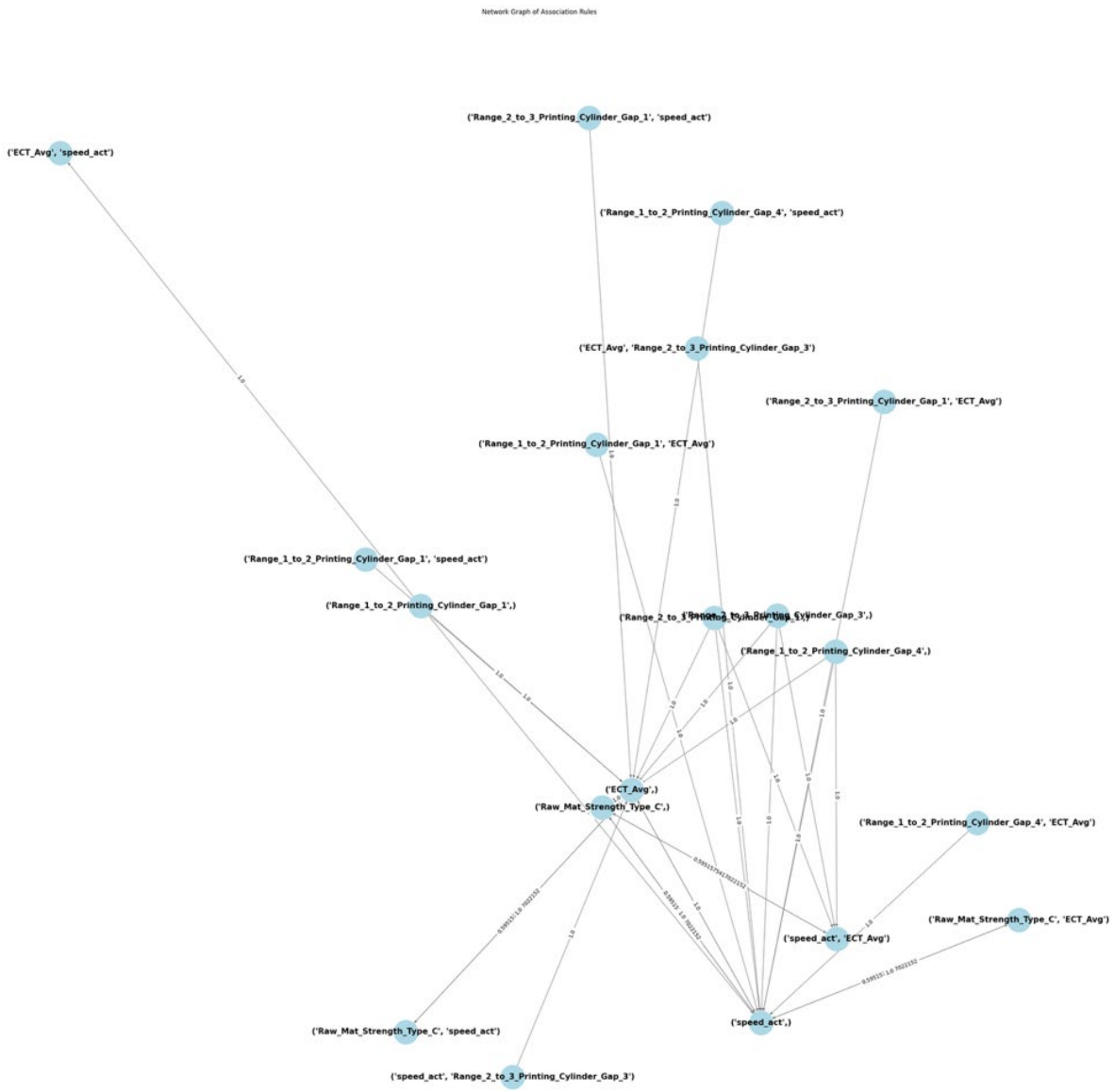
In the graph for results with 2 printing cylinder has a particular itemset has support =1.0. As the itemsets become more complex or less frequent, the support values decrease gradually with support values around 0.4-0.5. Several itemset show very low support values, this indicates these combinations are rare.



The results for 3 printing cylinders and 4 printing cylinders are the same. The top three patterns have support =1.0. Several patterns have support values around 0.603 and they occur frequently. A few patterns exhibit lower support values around 0.521.

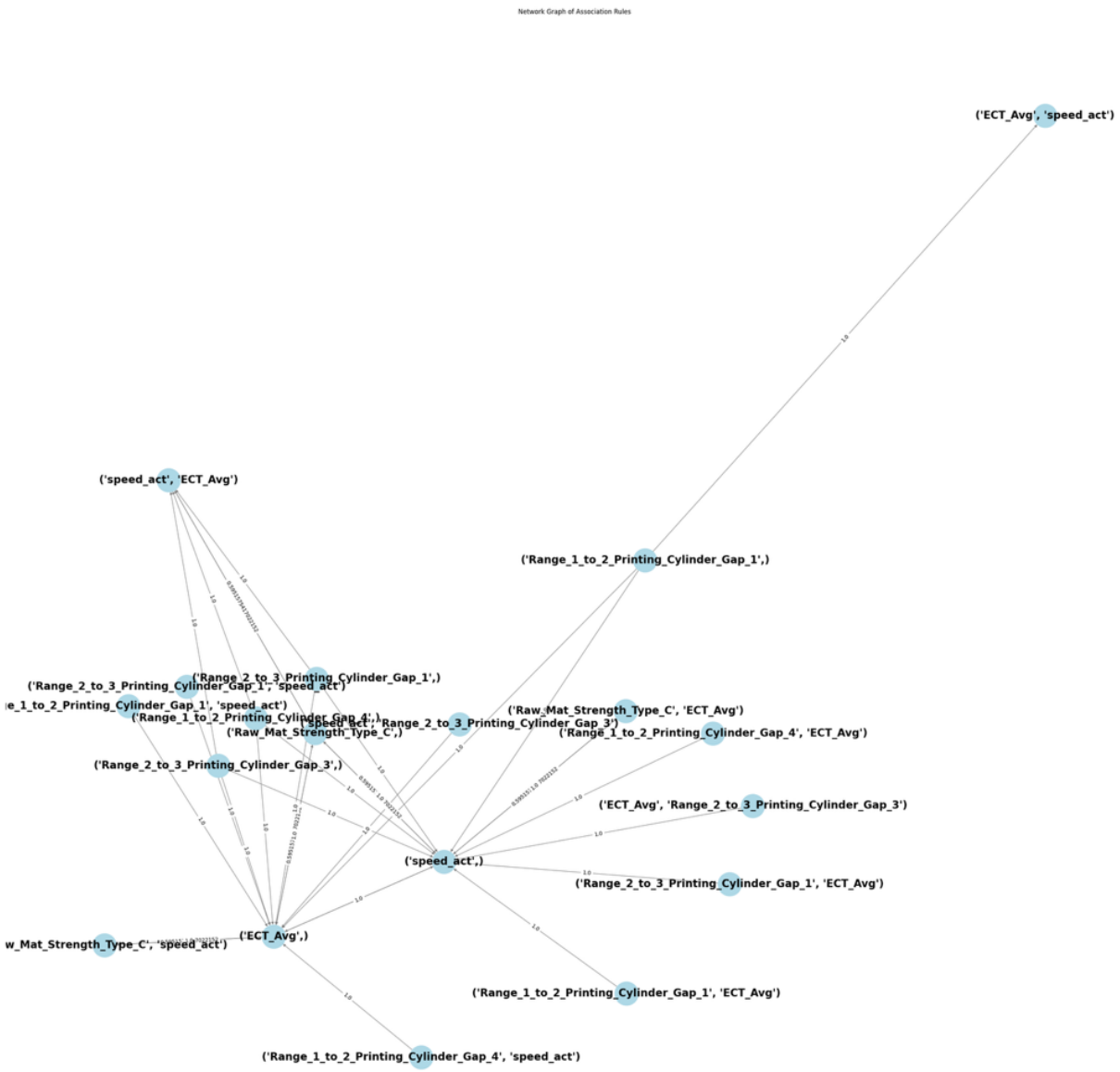
Association Rules

To understand the relationship between these variables, we generated network graphs that display associations between the frequent itemsets discovered through the apriori and fp-growth algorithms.



The graph focuses on the relationship between the ECT_Avg, speed_act and different printing cylinder gap ranges. It demonstrates how these variables are interrelated and affect each other in terms of product strength.

The graph displays more detailed associations between printing cylinders gaps, feed roll gaps and different raw material types and additives. These associations suggest important combinations that should be closely monitored during production.



In this network graph, interactions between various printing cylinder gaps, raw material strength and additive types are depicted. It helps to visualize how different combinations influence product outcomes.

Findings

For example, from the dataset of products produced using three printing cylinders, the FP-growth and Apriori algorithms suggest the following:

1. Controlling the printing cylinder gap of each printing cylinder unit within 0.0-1.0 and the average of all printing units should be between: 1.0-2.0.
2. Controlling the feed roll gap between: 1.0-2.0

3. Raw material type: C.
4. Incorporating additive type: G3.
5. Raw material strength: > 1.98

To verify the suggested result that the five variables affect the final product according to box production fundamentals (ISO Standard), we will select the sample data from product SKU "1397-123-00G3" (the top-selling product) following the five specified data pattern conditions suggested by the FP-growth and Apriori algorithms. We will then use a scatter plot to analyze the trend of BCT_Avg. It can be inferred that the pattern data of the printing cylinder and feed roll gaps applied to the raw material surface, raw material type, additive type, and raw material strength all have an impact on the final product strength (BCT_Avg), as observed in the actual production results.

Specifically, when the gap between the printing cylinders is low (indicating high pressure), the BCT_Avg is low. Conversely, when the gap is high (indicating low pressure), the BCT_Avg is high. This suggests that controlling the printing cylinder gaps can help optimize the product strength. Additionally, the raw material type, additive type, and raw material strength also influence the BCT_Avg, and these variables should be considered when developing the production recipe.

By applying these condition patterns and analyzing the scatter plot, we can gain insights into how different combinations of variables affect the final product strength.

Data Clustering

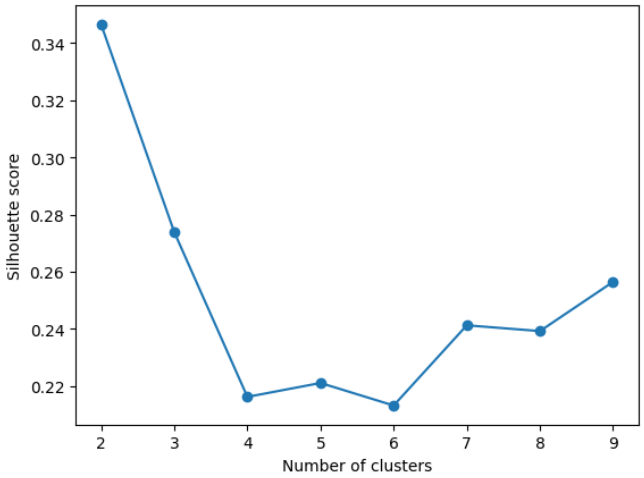
Frequent item mining provided valuable insights on the parameters that directly influence the strength of the final product. However, there is more that can be learned from these variables to determine if a box will pass the strength test. The goal of this section is to determine the relationship that the average strength of the final product (BCT_Avg) has on factors such as the gap distance between feeding rollers, gap distance between printing cylinders, and the average strength of raw materials that the box is made of. It should be noted that manufacturing specifications for each box produced varies. This will dictate the number of printing rollers that are used to create the final product.

This section will apply data clustering techniques, such as K-Means, Ward Hierarchical Clustering, and Spectral Co-Clustering, to obtain a visualization of the relationships. These algorithms allow for a specified number of clusters to be designated, and although they accomplish the same objective, the way each algorithm achieves it is different. For example, using the K-Means algorithm executes a clustering visualization more quickly than Ward Hierarchical or Spectral Co-Clustering algorithms.

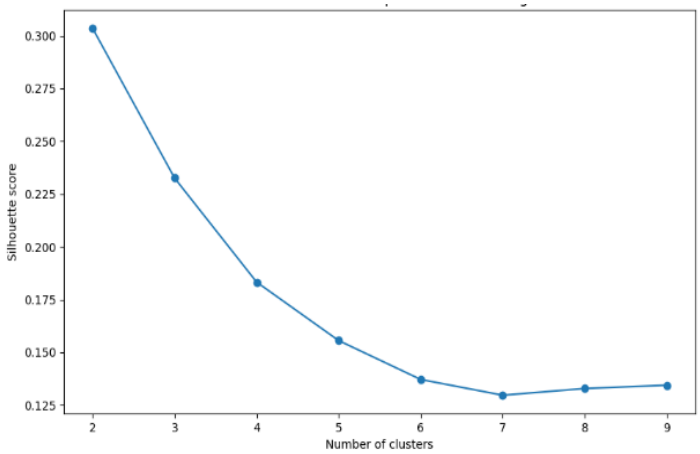
The first step towards creating cluster visualization is determining the ideal number of clusters. Using the Silhouette score approach, the appropriate number of clusters can be

identified. The chart below shows the ideal clustering score for the datasets that deal with products requires one and three printing cylinders.

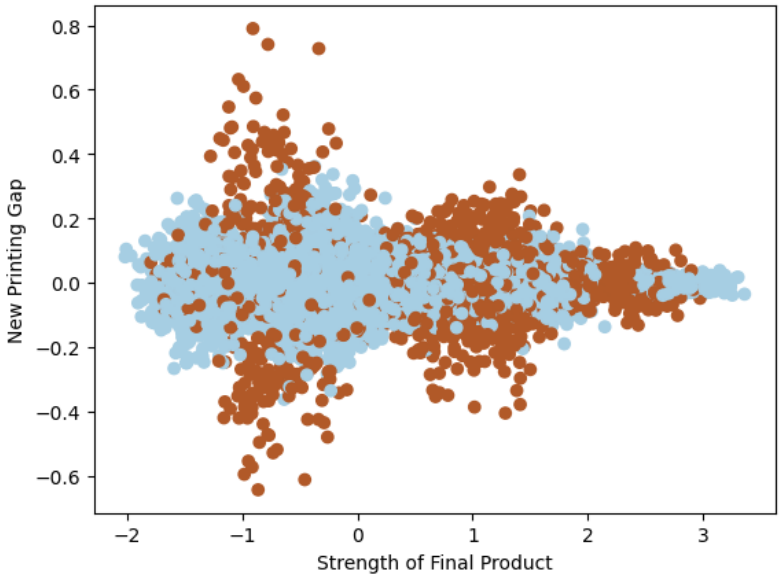
1 Printing Cylinder



3 Printing Cylinder

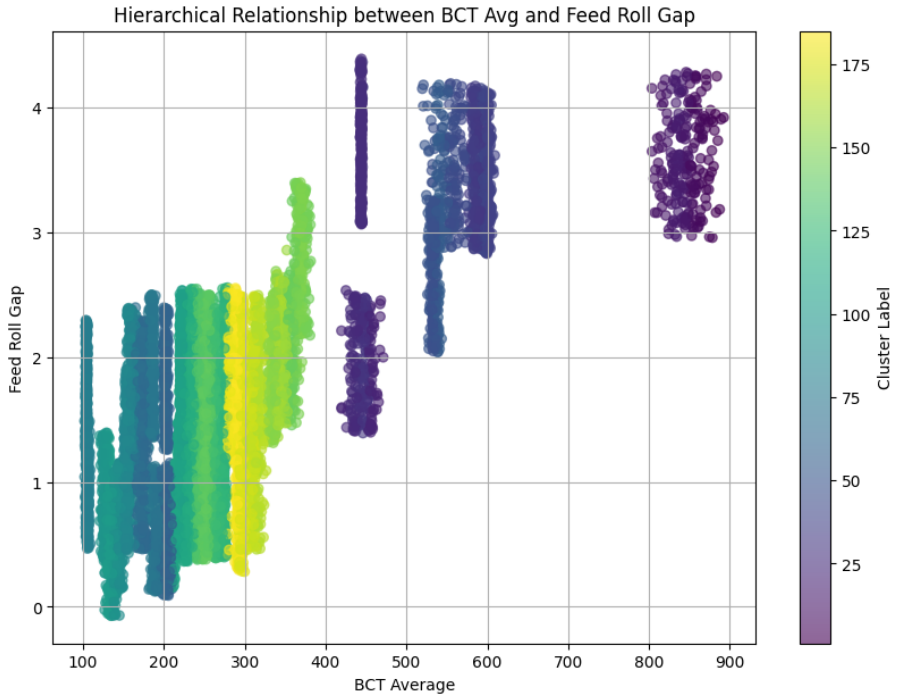


Although the shape of each graph is different, both Silhouette scores indicate that the ideal number of clusters, regardless of the number of printing cylinders used is two when K-Means clustering is applied. Additionally, Ward Hierarchical and Spectral Co-Clustering also received a Silhouette score of two. When executed, these algorithms provided different visualizations based on how each algorithm clusters the data.



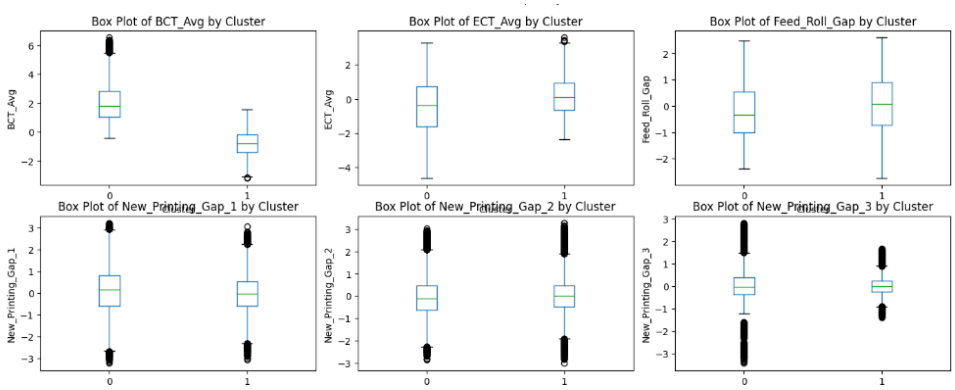
The image above depicts K-Means clustering applied to the one printing cylinder gap dataset. It can be determined from this graph that the ideal printing gap size is between 0.0 and 0.2 for the final product to retain its structural integrity. Ward Hierarchical clustering was helpful in identifying outliers and seeing how clusters were merged. Additionally, further analysis was

able to show the hierarchical relationship between parameters, such as the relationship between BCT average and Feed Roll Gap. As seen in this graph, as the BCT average increases, so does the feed roll gap size. This indicates that the pressure applied to the final product by the feed rollers significantly impacts how strong the final product will be.



Other methods to view the data distribution and outliers were applied to the datasets. For instance, a box plot graph was created based on the output from the Spectral Co-Clustering approach that resulted in the input below.

Spectral Co-Clustering Box Plot



The conclusion using this approach is that the data distribution appears to be normal within each cluster. However, there were some apparent outliers.

In addition to box plots, heat maps were also created. To create the heat maps, the average of each variable in each cluster was calculated and plotted. The heat maps of all three clustering methods can be correctly interpreted following the corrugated box strength calculation principle.

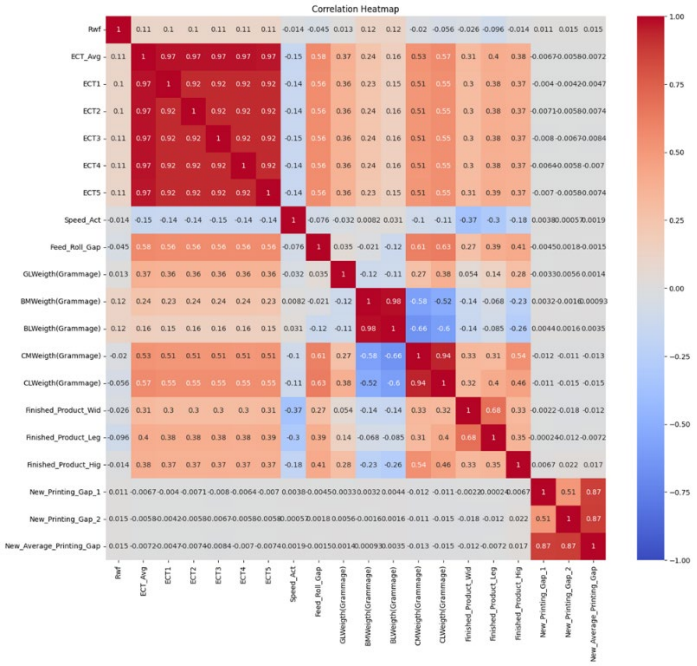
Using the three printing cylinder dataset as an example, the Ward Hierarchical Clustering Heat Map, indicated that although the raw material strength average in cluster 1 is lower than in cluster 0, the feed roll and printing gaps are less applied on the raw material surface. It can be concluded that because the product specifications indicate that multiple printing cylinders must be used, less pressure is applied to the raw material as it moves through the feed roller and first printing cylinder.

The conclusion from applying clustering methods to the datasets is that the gap distance between the feed rollers and printing cylinders varies depending on the box specifications. If the box requires only one printing cylinder, the gap can be slightly smaller between the feeding rolls and printing cylinders. This should not greatly affect the strength of the final product. However, if a box must use more than one printing cylinder, the pressure should be distributed so the structural integrity of the box remains intact.

Dimensionality Reduction

It is a technique for reducing the number of features in a dataset while preserving most of the information to better visualize the data. It maps higher dimensional data to a lower dimension. We apply different dimensionality reduction algorithms to our dataset to compare the results.

Correlation of features



A heatmap is generated to plot the correlation between the dataset features. PCA reduces redundancies by generating a new set of principal components which are orthogonal to each other.

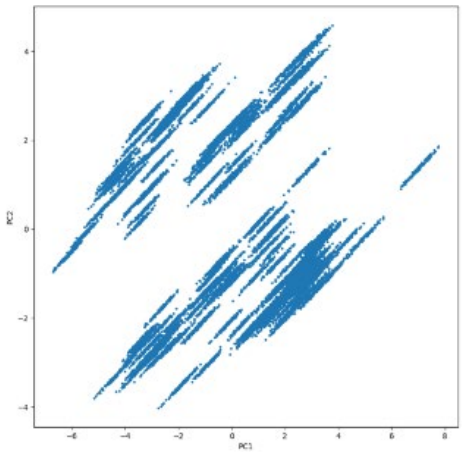
From the heatmap, we can see that the dataset features are highly correlated. ECT1 to ECT5 have correlations above 0.9. Similarly, the new_printing_gap 1 and 2 are highly correlated to the new_printing gap average. PCA can reduce the feature space by combining them into a principal component that represents shared variance between the variables.

Principal Component Analysis (PCA)

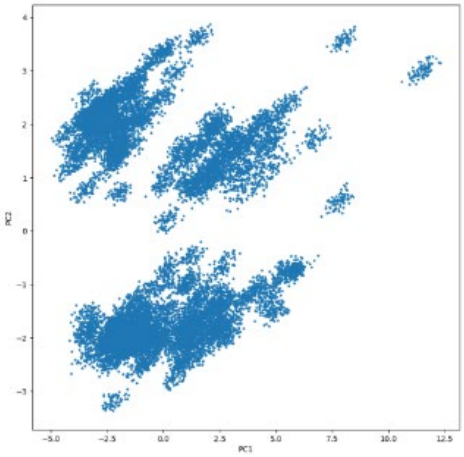
It is a technique used to reduce the dimensions of large data sets. by transforming a large set of variables into a smaller one that still contains most of the information in the large set. It is a linear dimensionality reduction technique; therefore, it does not identify any non-linear patterns in the data.

PCA is applied to five datasets with different number of printing cylinders used and the results are recorded. The graphs are generated for PCA as follows.

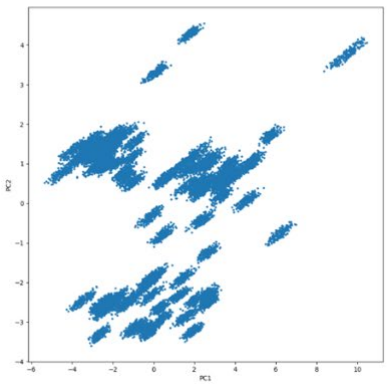
0 printing cylinders



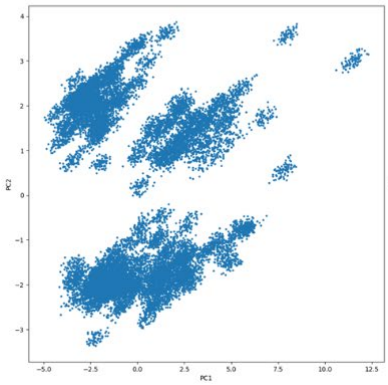
1 printing cylinder



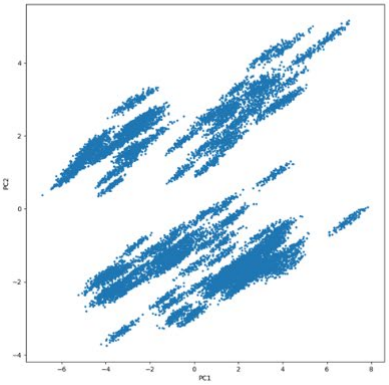
2 printing cylinders



3 printing cylinders



4 printing cylinders



In terms of overall shape, the graphs are similar. Graphs for printing cylinders 1,2, and 3 are almost identical. Printing cylinders 1 and 4 have sharper edges for the clusters. The clusters are well separated and spread out

Number of printing cylinders used	Explained Variance Ratio
0	[0.39957935, 0.22999367]
1	[0.34686112, 0.24216504])
2	[0.38585775, 0.16069194]
3	[0.33846826, 0.17175637]
4	[0.38321679, 0.1881126]

For each of the five datasets, the explained variance ratio for PCA is generated. As visible from the table above, the total variance does not exceed 60%. This is very low and indicates that PCA is not doing a great job of capturing some aspects of the dataset. Since PCA only identifies linear data, we try kernel PCA next to see if any non-linear trends are visualized.

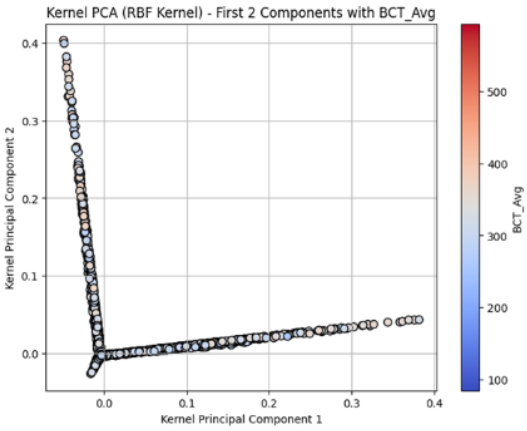
Kernel PCA

Kernel Principal Component Analysis (Kernel PCA) is an extension of traditional PCA that allows for the identification of principal components in data with complex, non-linear relationships. By applying a kernel function, Kernel PCA maps the original data into a higher-dimensional space where linear separations are possible, enabling more effective dimensionality reduction for non-linear datasets.

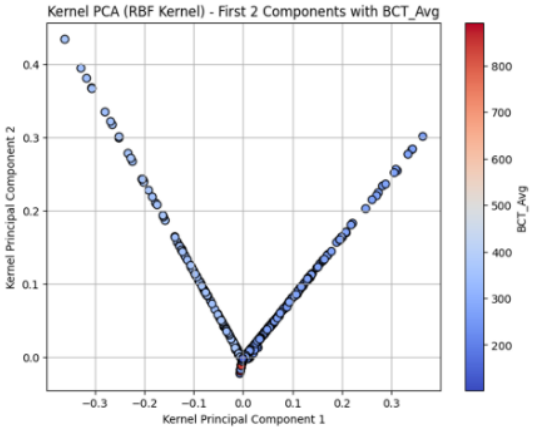
3 kernel functions were implemented. They are RBF, Sigmoid, and Poly.

1) **RBF (Radial Basis Function) Kernel**: The RBF kernel is one of the most used in Kernel PCA, as it captures local relationships in the data by mapping points based on their distance from each other. It is well-suited for datasets with non-linear and intricate patterns.

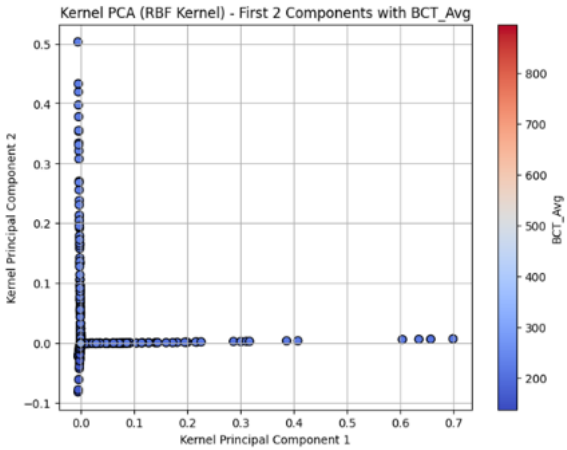
0 printing cylinders



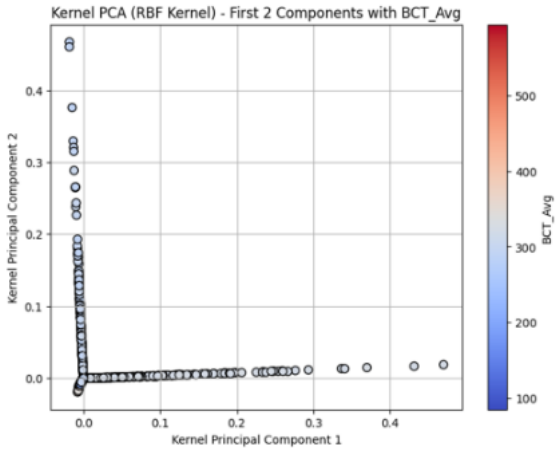
1 printing cylinder



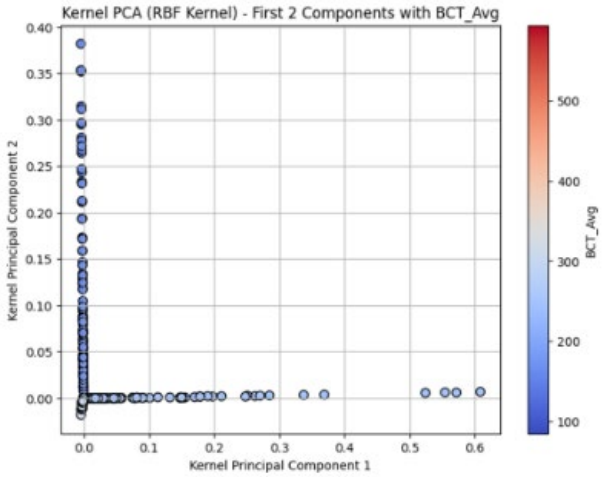
2 printing cylinders



3 printing cylinders



4 printing cylinders



The graphs show PCA1 and PCA2 plotted against the target variable "BCT_Avg," which represents the final strength of the corrugated box. Upon reviewing the graphs, it is evident that they are quite similar, with the exception of printing cylinder 1, where the data points increase as PC1 values increase. Looking at the graph for RBF, it is apparent that the points are evenly distributed for both principal components.

Number of printing cylinders used	Explained Variance Ratio (RBF)	Explained Variance Ratio (Sigmoid)	Explained Variance Ratio (Poly)
0	[0.5311204 0.4688796]	[0.58727972 0.41272028]	[0.75943827 0.24056173]
1	[0.5026503 0.4973497]	[0.5718121 0.4281879]	[0.84410681 0.15589319]
2	[0.52094482 0.47905518]	[0.67184547 0.32815453]	[0.90945889 0.09054111]
3	[0.51641813 0.48358187]	[0.6404411 0.3595589]	[0.57549561 0.42450439]
4	[0.5199131 0.4800869]	[0.65552626 0.34447374]	[0.63094804 0.36905196]

The table displays the explained variance of each printing cylinder for every kernel used. The total variance captured by kernel PCA is approximately 99%, a substantial improvement from the 60% variance captured by regular PCA. This indicates significant non-linearity in the data, making Kernel PCA more suitable for this analysis.

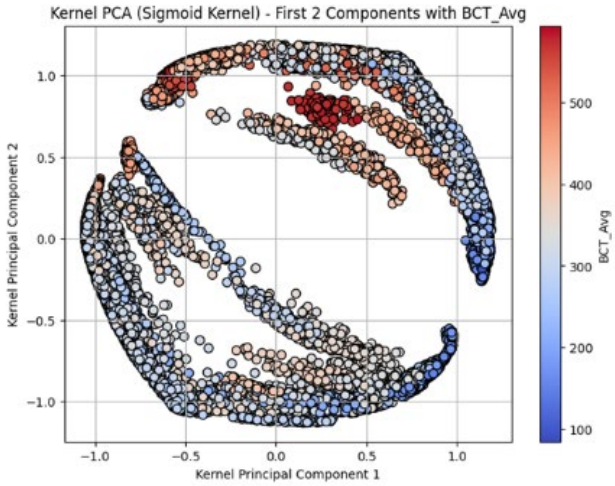
In all the RBF graphs, there is no distinction between the clusters formed and the value of BCT_Avg (Target variable).

Some interesting observations:

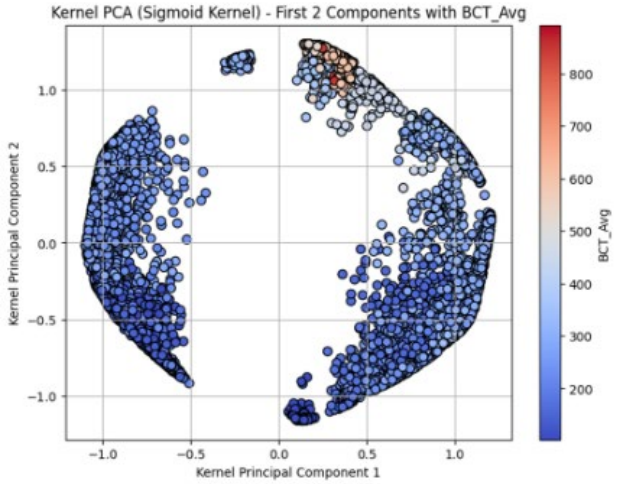
- On the whole, RBF has the most balance between PC1 and PC2 in terms of the variance captured. For sigmoid, the difference between PC1 and PC2 is more than RBF.
- Polynomial has the most disparity, with PC1 capturing most of the variance.
- For Sigmoid and Polynomial, the variance captured by PC1 seems to follow a curve where it increases from printing cylinders 0 to 2 and then decreases.

2) **Sigmoid Kernel**: The sigmoid kernel, inspired by neural networks, transforms data using the hyperbolic tangent function. It models pairwise interactions between data points, but its performance can be sensitive to the choice of parameters.

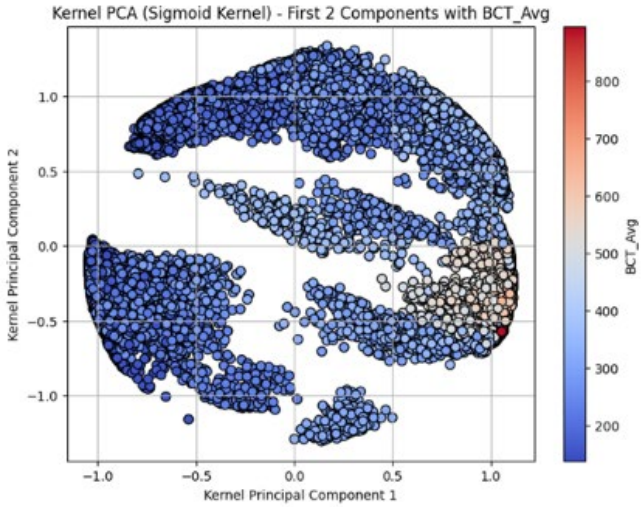
0 printing cylinders



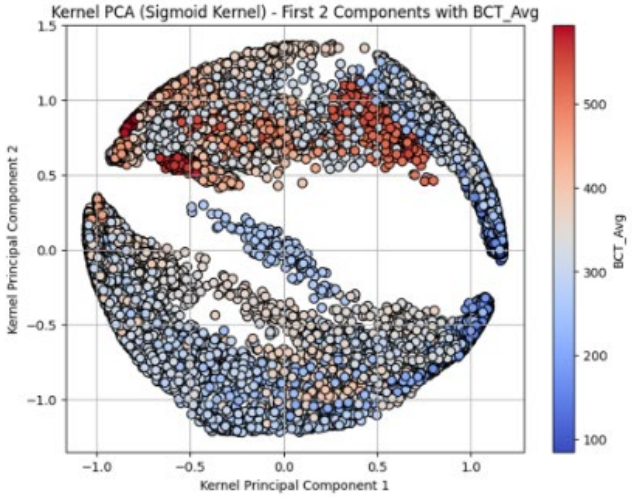
1 printing cylinder



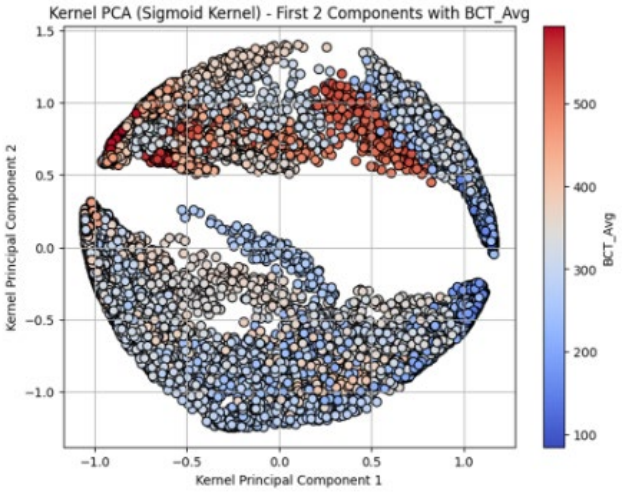
2 printing cylinders



3 printing cylinders



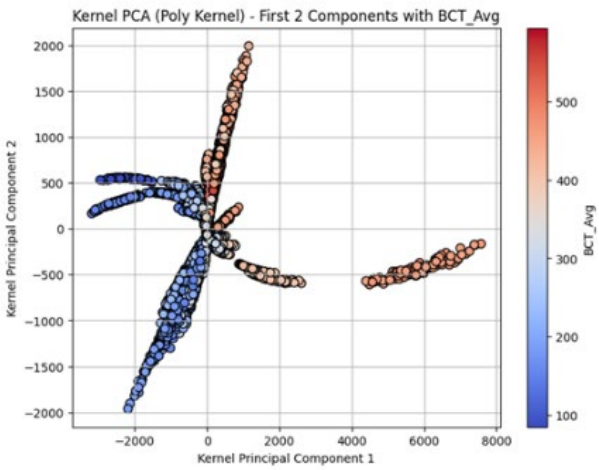
4 printing cylinders



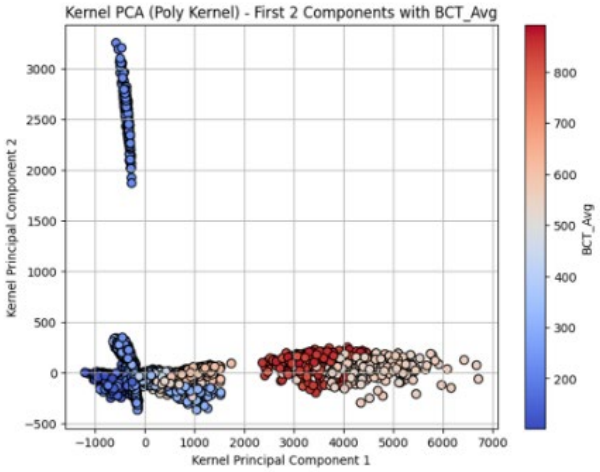
All the graphs have a spherical shape. The sigmoid plots have more separation between the different values of BCT_Avg with similar BCT_Avg values clustered together.

3) **Polynomial Kernel**: The polynomial kernel extends the standard dot product in PCA by raising it to a power, enabling the identification of polynomial relationships between features. It can capture more complex data patterns but may be less effective for highly non-linear data compared to RBF.

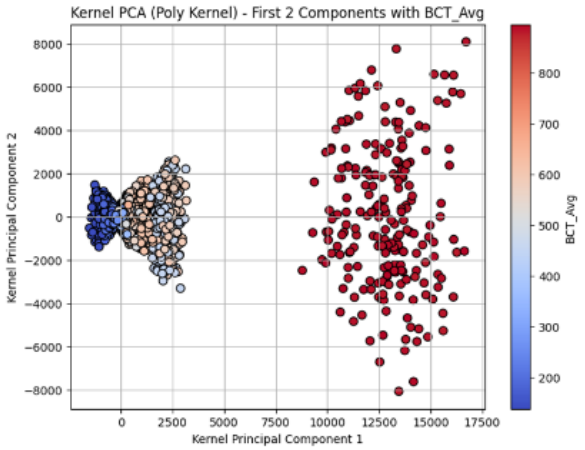
0 printing cylinders



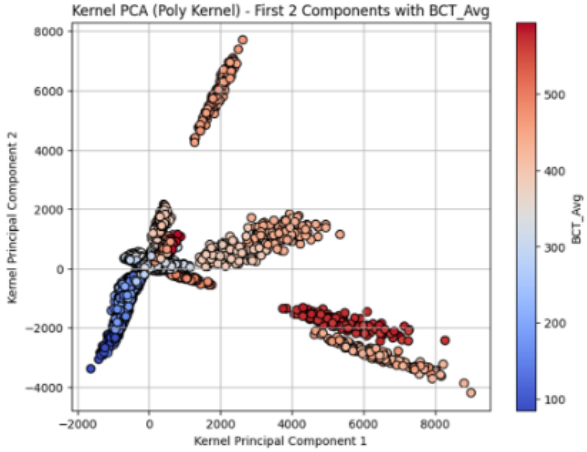
1 printing cylinder



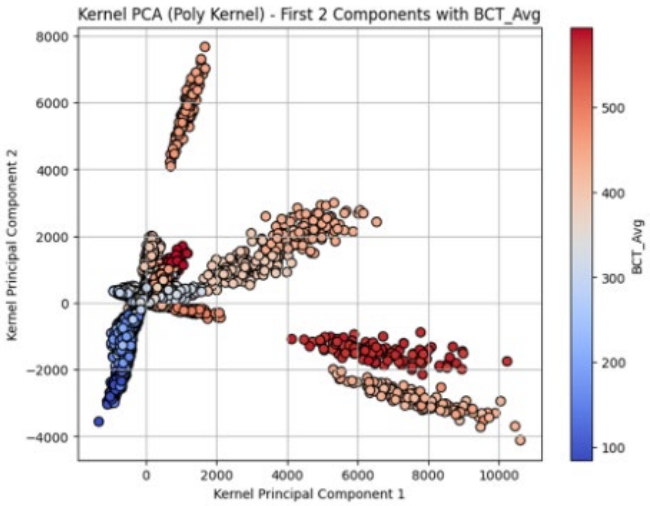
2 printing cylinders



3 printing cylinders



4 printing cylinders



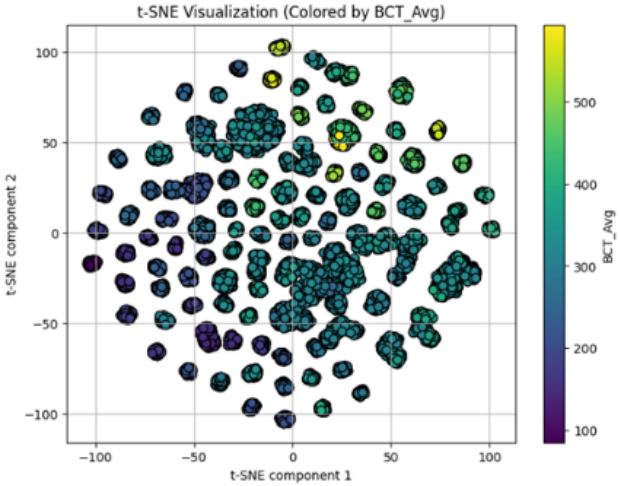
The polynomial plots give the most amount of distinction between different values of BCT_Avg, with similar values being in the same region of clusters. The graphs, except for printing cylinder 2, all have similar patterns. Printing cylinder 2 has a very unique pattern. An interesting thing to observe is that the red points are scattered across for printing cylinder 2 where for the rest they are tightly packed together.

t-SNE (t-distributed Stochastic Neighbor Embedding)

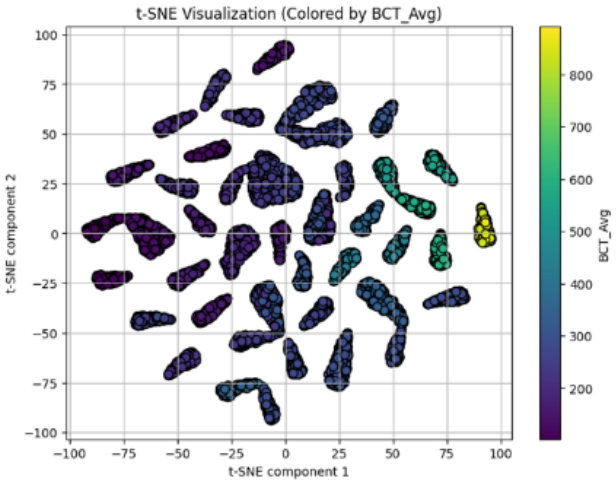
It is a popular dimensionality reduction technique used for visualizing high-dimensional data. It focuses on preserving the local structure of the data by minimizing the divergence between two probability distributions that represent pairwise similarities in both the high-dimensional and low-dimensional space. t-SNE is widely used for clustering and visualizing non-linear structures but

can sometimes be computationally intensive and is sensitive to parameter tuning, such as perplexity.

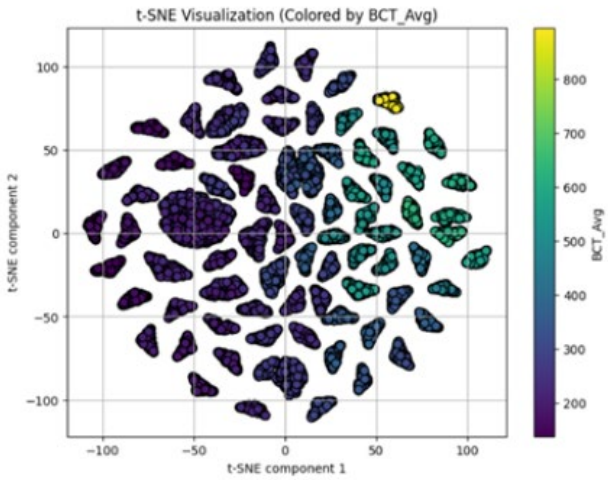
0 printing cylinders



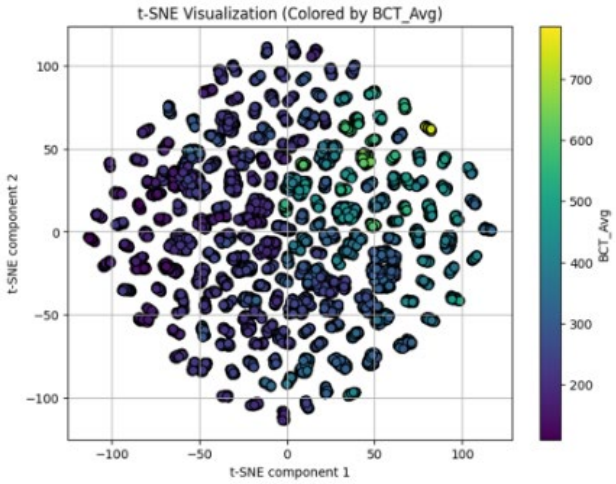
1 printing cylinder

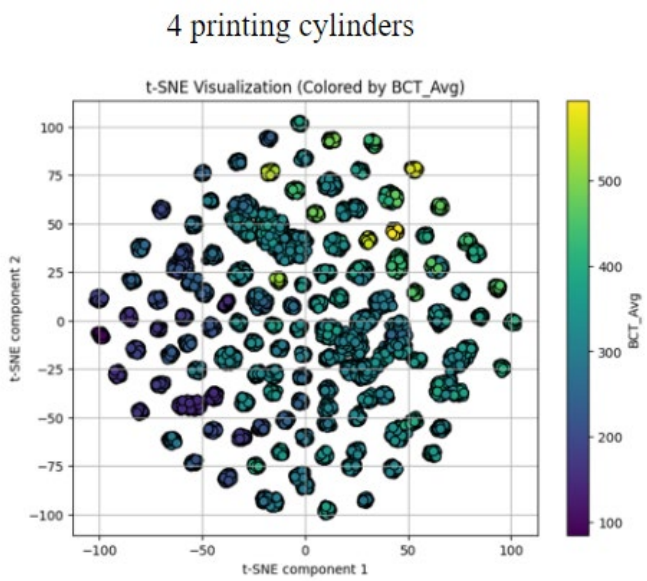


2 printing cylinders



3 printing cylinders



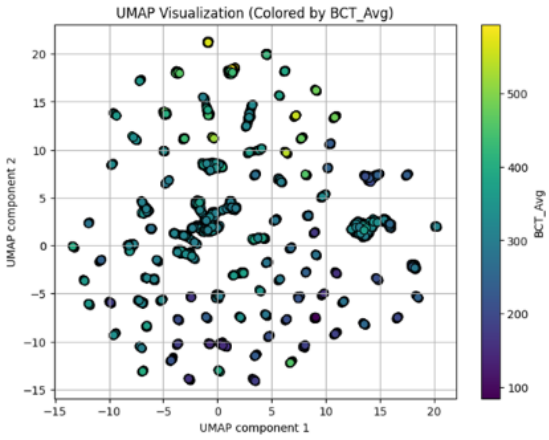


Like poly kernel clustering, t-SNE too has similar values to BCT_Avg clustered together. All the graphs for t-SNE are similar in terms of the structure and values spread. The 0-printing cylinder plot has captured mostly the lower end of the BCT_Avg values.

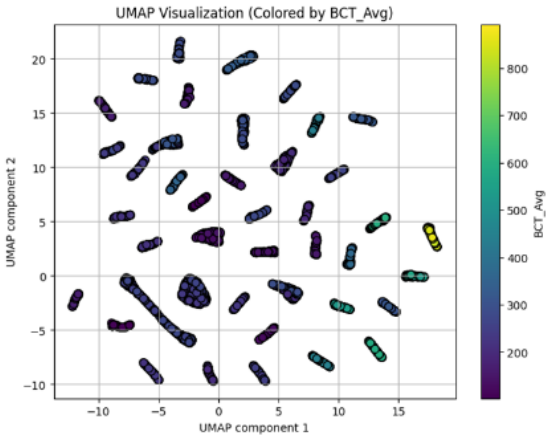
UMAP (Uniform Manifold Approximation and Projection)

UMAP is another dimensionality reduction method that emphasizes both local and global structure in the data. It operates on the assumption that the data lies on a low-dimensional manifold, capturing non-linear relationships effectively. UMAP is often faster than t-SNE, can maintain a more global structure, and is scalable to larger datasets, making it suitable for visualization and clustering of complex data.

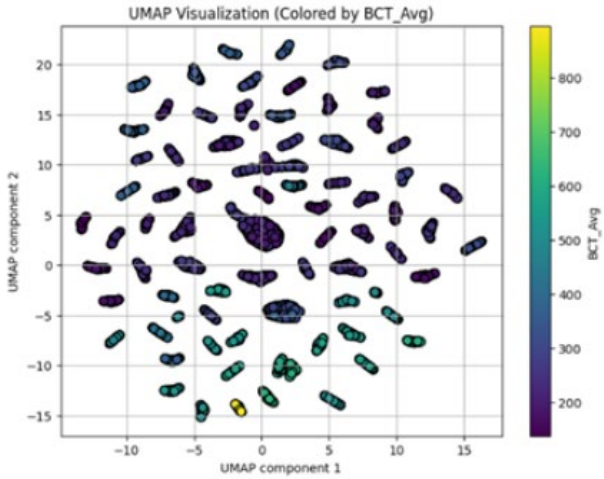
0 printing cylinders



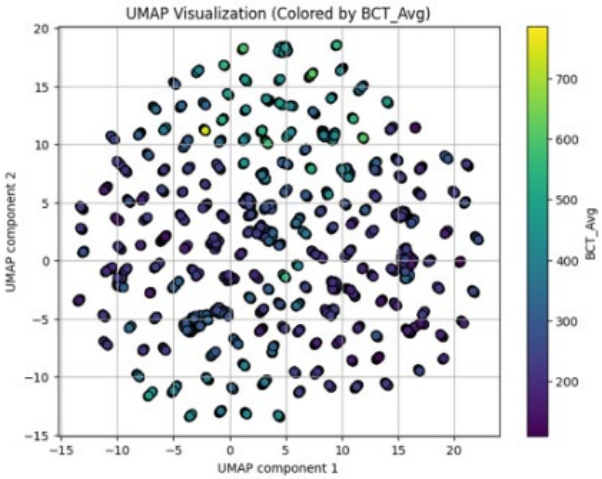
1 printing cylinder



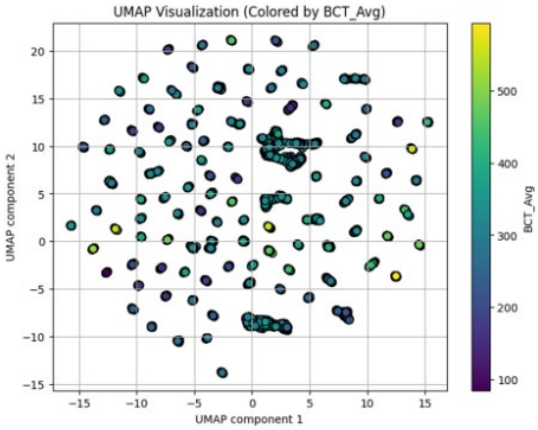
2 printing cylinders



3 printing cylinders



4 printing cylinders



UMAP and t-SNE have different ways of clustering data. UMAP doesn't always separate clusters based on the BCT_Avg value, and there can be overlap between different BCT_Avg values within clusters. This is because UMAP considers the global structure of the data. On the other hand, t-SNE focuses on the local structure of the data, which might allow it to distinguish points better.

Supervised Learning

This section on supervised learning will apply logistic regression techniques on all datasets to prevent overfitting and determine the prediction accuracy. Additionally, the techniques on each dataset have been broken down. The model will start making predictions on the strength of the final product using data from products that are manufactured using no printing cylinders. The project team will iterate through each number of printing cylinders used during production to see how accurate the model is.

To get started, explanatory variables must be identified. The research conducted in earlier sections identified raw material strength (ECT), machine speed, and the distance between cylinders and feed rollers as parameters that work best as explanatory variables. These variables influence the strength of the final product (BCT), which is our target variable. If BCT is equal to or exceeds the average strength of a known product, then the box will pass the strength test.

Zero Printing Cylinders

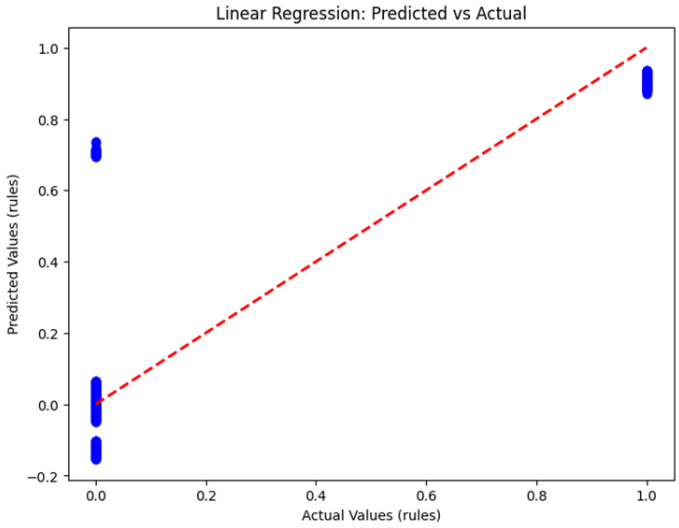
```
Accuracy: 1.0
Confusion Matrix:
 [[9373    0]
 [   0  825]]
ROC-AUC Score: 1.0
RMSE for Logistic Regression: 0.005379793705857176
Pseudo R-Squared (McFadden): 0.9999105827018364

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      9373
           1       1.00      1.00      1.00       825

    accuracy                           1.00     10198
   macro avg       1.00      1.00      1.00     10198
weighted avg       1.00      1.00      1.00     10198

Mean Squared Error for Linear Regression: 0.0086469555660127
R-squared for Linear Regression: 0.883705098023182
```

The results show an accuracy of 1, this indicates the model is predicting the class for every instance correctly in the test set. The confusion matrix shows perfect classification with no false positives or false negatives. All 9373 non-rejected items were correctly classified as 0 and 825 rejected items as 1. The ROC-AUC score shows the model is suitable for distinguishing between rejected and non-rejected products.

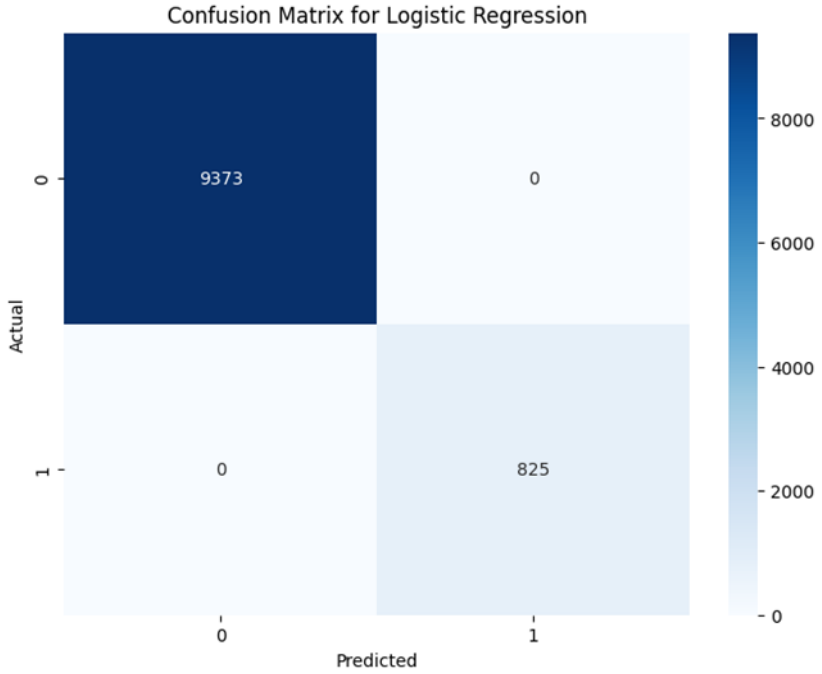
The RMSE is very low, in this case, the probabilities predicted are close to the true value. The Pseudo R-squared is close to 1, which means the logistic regression model fits the dataset.

Precision, recall and F1-score are all 1, this confirms that the model performs perfectly in both classes. The Mean Squared Error is 0.0086, indicating the average squared difference between predicted and true values.

The linear regression model below explains the 88.37% R-squared shows the variance in the rejection status.



The plot shows the predictions for non-rejected and rejected cluster well around their respective values. However, it is clear that logistic regression performs better.



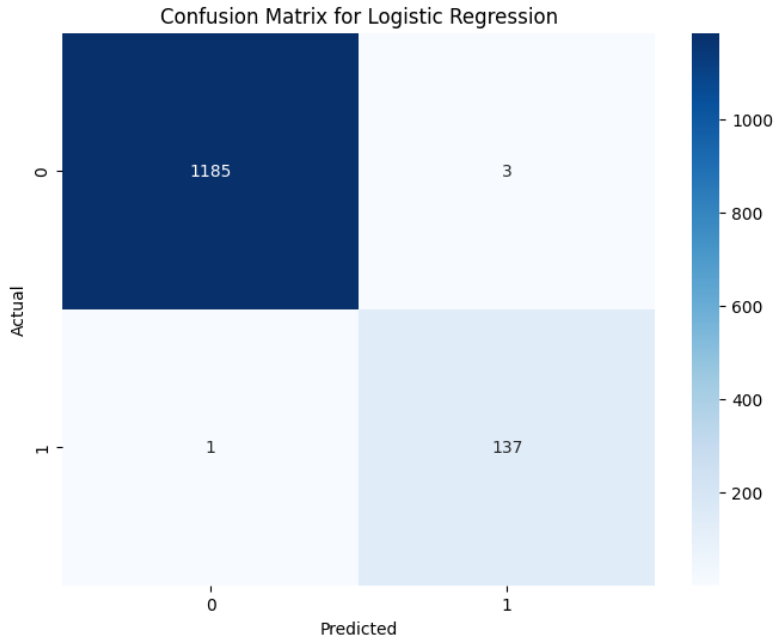
The confusion matrix for logistic regression shows perfect classification, aligning with the overall evaluation metrics.

One Printing Cylinder

Supervised learning on this dataset begins by splitting the data into training and test sets and initializing the model. Logistic regression is also applied so that overfitting does not occur in the model. Once these steps are completed the data is split into training and test data. The model will learn from the training data and the test data will be introduced to see how accurate the model's predictions are.

```
Accuracy: 0.9969834087481146
Confusion Matrix:
 [[1185    3]
 [   1  137]]
ROC-AUC Score: 0.9951141853315767

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1188
           1       0.98      0.99      0.99       138

    accuracy                           1.00      1326
   macro avg       0.99      1.00      0.99      1326
weighted avg       1.00      1.00      1.00      1326
```

```
Mean Squared Error for Linear Regression: 0.054004786050283186
R-squared for Linear Regression: 0.4208063778037151
```



The results of the accuracy scores are encouraging. All metrics indicate a model that is almost perfect at predicting if the structural integrity of a box will pass the strength test. Because the margin of error is small, it means that this model can be used during the manufacturing process to significantly reduce the number of times the production line must be shut down. This will help to mitigate financial losses incurred by such events and improve productivity in the warehouse. It should be noted that this dataset contains information related to boxes that are produced using

one printing cylinder. The next sections of this report will apply this logistical model to datasets with information on items that are manufactured using two or more printing cylinders.

Result and Conclusion of 2 Printing Cylinder Gaps:

Linear regression results:

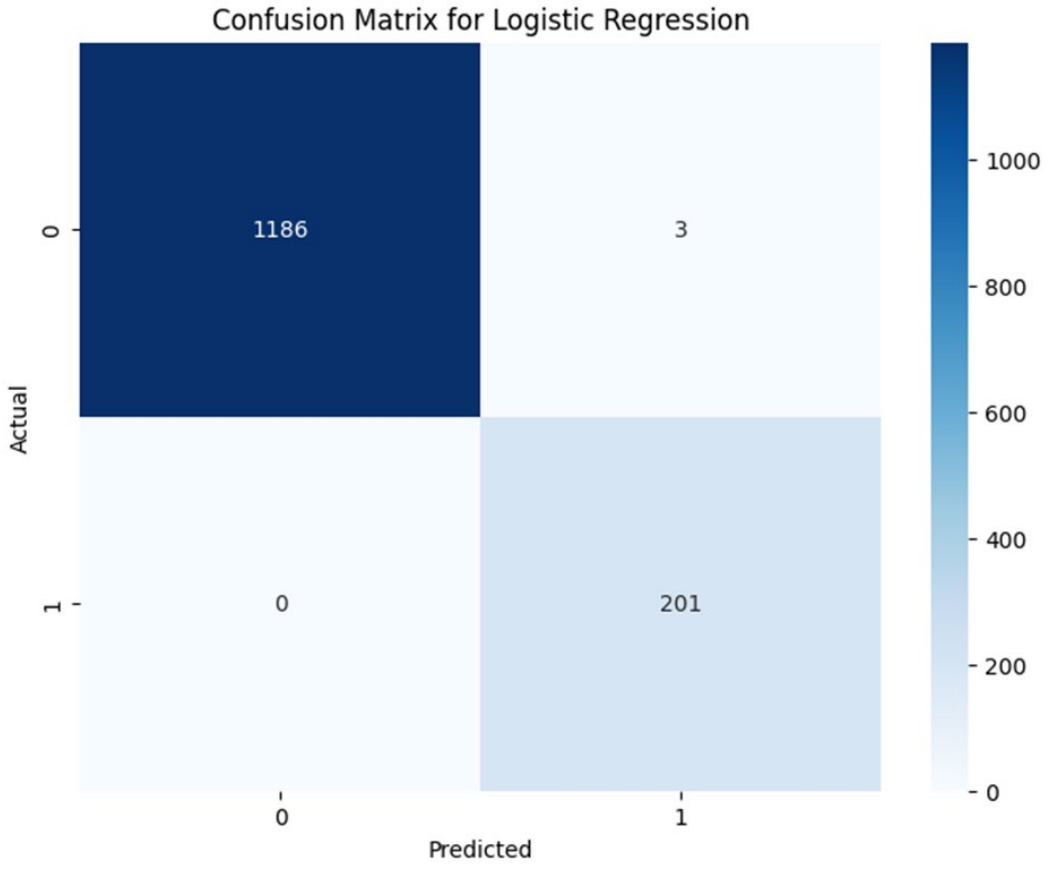
```
Mean Squared Error for Linear Regression: 0.063413403075568
R-squared for Linear Regression: 0.48733608625374003
```

The R-squared value of 0.487 suggests that approximately 48.7% of the variability in the actual data is explained by the linear regression model. This indicates a moderate level of explanatory power, meaning the model captures some, but not all, of the underlying patterns in the data. There might be other important predictors not included in the model, or the relationship between predictors and the target might not be linear.

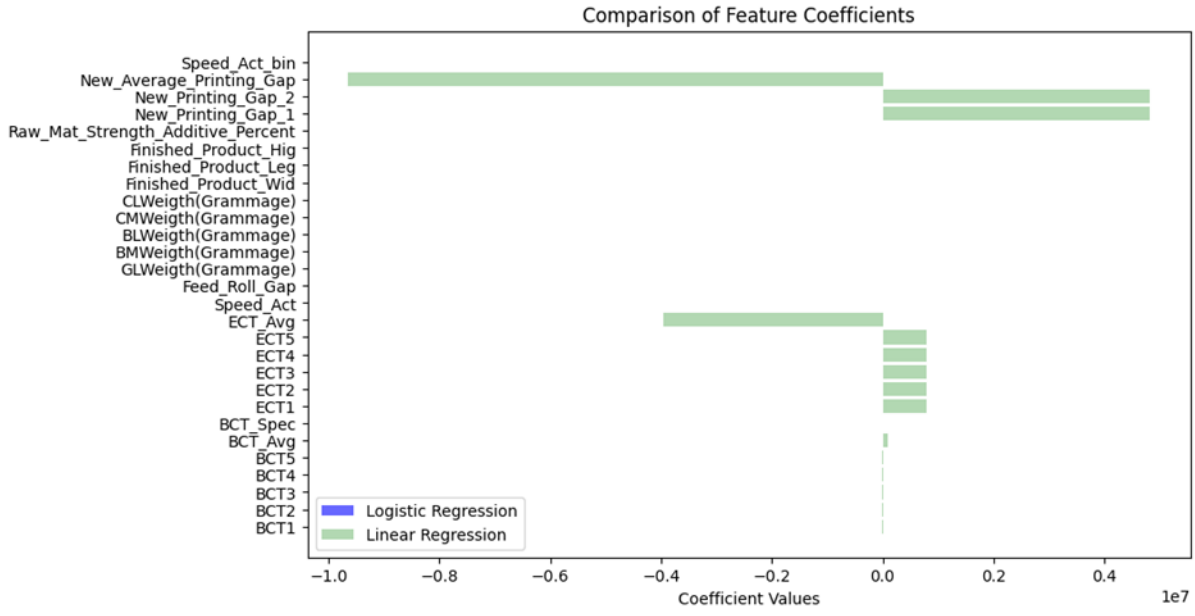
Logistic regression results:

```
Accuracy: 0.9978417266187051
Confusion Matrix:
 [[1186    3]
 [   0  201]]
ROC-AUC Score: 0.9987384356602187

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1189
           1       0.99      1.00      0.99       201

    accuracy                           1.00      1390
   macro avg       0.99      1.00      1.00      1390
weighted avg       1.00      1.00      1.00      1390
```

Many metrics for classification are considered here. Both accuracy and ROC score are 99% indicating that the model can predict the classes well.



The confusion matrix for the logistic regression model is plotted. We can see that the logistic regression is a much better fit for the data.



The above bar chart compares feature coefficients from two models: logistic regression (in blue) and linear regression (in green). The x-axis represents the coefficient values, while the y-axis lists the features. The features "New_Average_Printing_Gap_2" and "New_Printing_Gap_1" have significant positive coefficients in the linear regression model, indicating a strong positive relationship with the target variable.

"Speed_Act_bin" has a large negative coefficient, suggesting an inverse relationship. The logistic regression coefficients are not visible, implying they might be negligible.

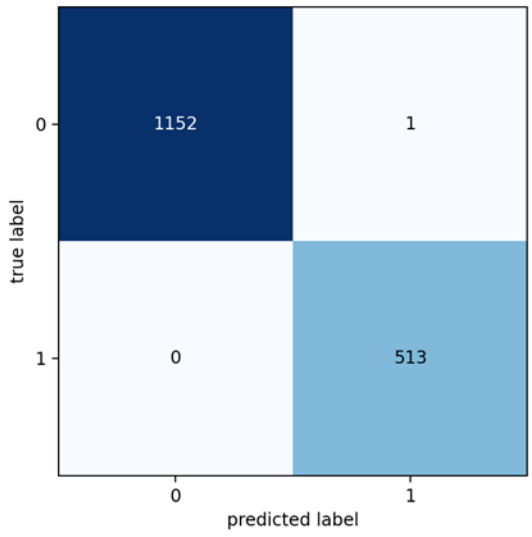
3 Printing Cylinder Gaps:

```
RMSE for Logistic Regression: 0.030996664899294677
Pseudo R-squared (McFadden) for Logistic Regression: 0.9996828995010857
Mean Squared Error for Linear Regression: 0.10613172072249702
R-squared for Linear Regression: 0.5019780008729647
```

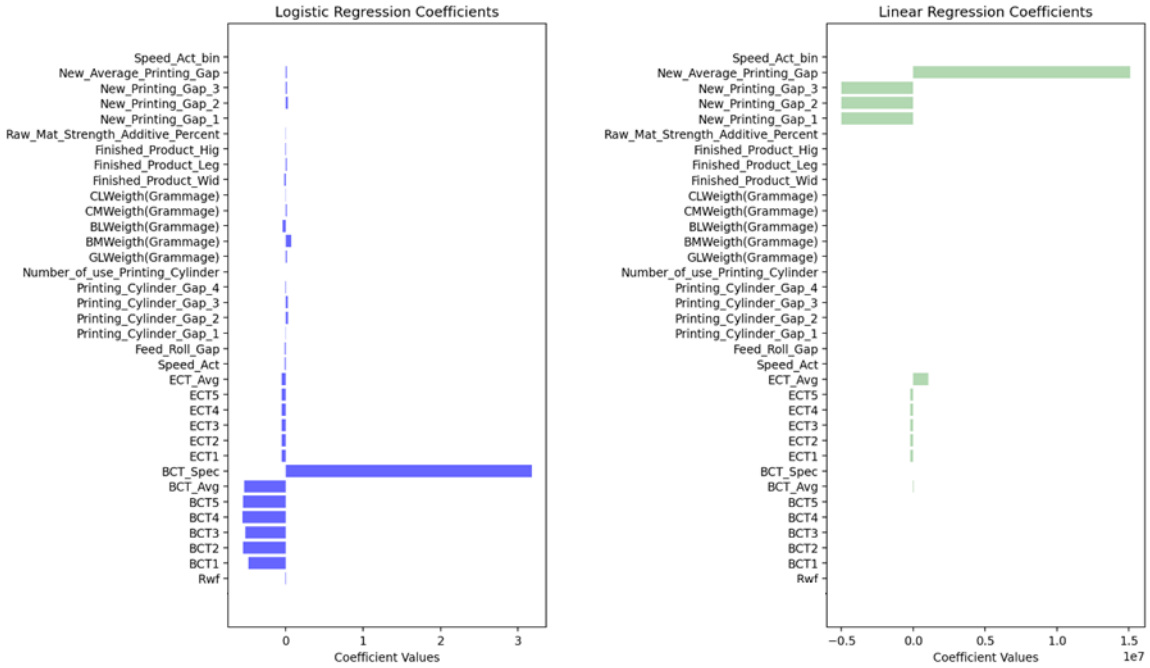
As a result of 3 printing cylinder gaps, it can be interpreted as

- **Logistic Regression** seems to perform exceptionally well, with a very low RMSE and almost perfect McFadden pseudo R-squared. This suggests the model is very well-fitted to the data.
- **Linear Regression** has a higher MSE and a moderate R-squared value, indicating that it does not fit the data as well as the Logistic Regression model.

This is a confusion matrix, commonly used to evaluate the performance of a classification model by using logistic regression.



- There is only 1 false positive, meaning the model very rarely predicts "1" incorrectly when it should predict "0" and there are no false negatives, meaning the model does not miss any true positives.



Logistic regression appears to involve a broader range of variables than linear regression, which likely contributes to its higher accuracy. However, it's important to note that the dominant features influencing the Logistic Regression model are related to BCT (Box Compression Test) values, while the Linear Regression model is primarily driven by features related to machine printing cylinder gaps. This distinction suggests that each model relies on different aspects of the data to make predictions, with Logistic Regression capturing more complex relationships across multiple variables.

4 Printing Cylinder Gaps:

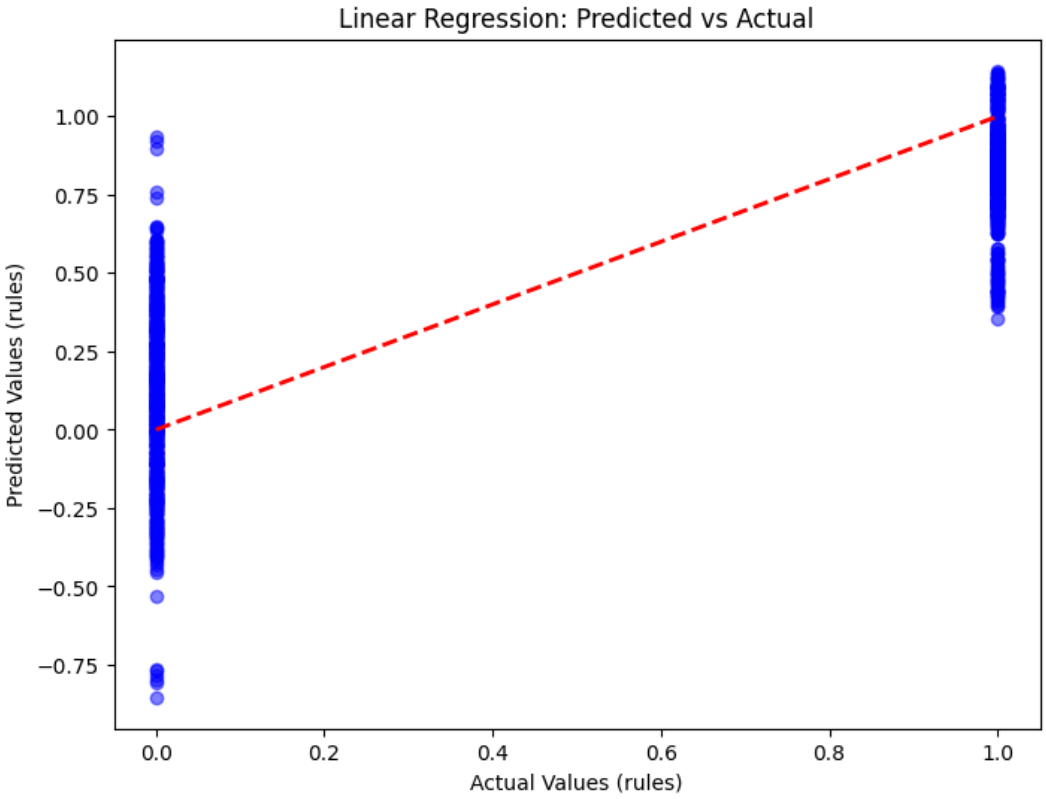
```
RMSE for Logistic Regression: 0.04440074543263458
Pseudo R-squared (McFadden) for Logistic Regression: 0.9995795744272703
```

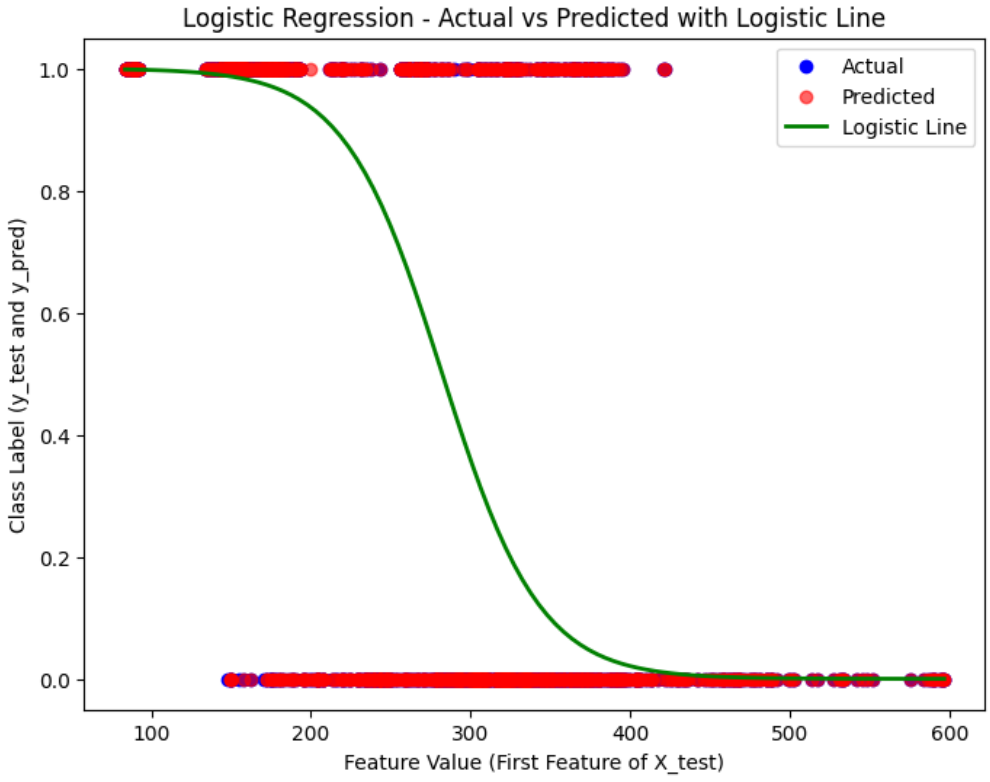
```
Mean Squared Error for Linear Regression: 0.07147767511243602
R-squared for Linear Regression: 0.7124997127402333
```

As a result of 4 printing cylinder gaps, it can be interpreted as

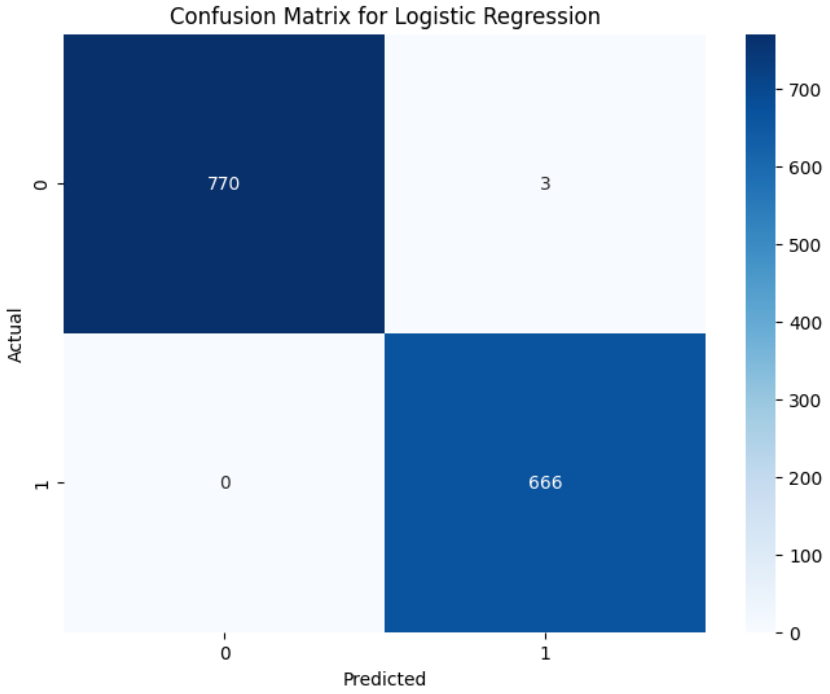
- **Logistic Regression** seems to perform exceptionally well, with a very low RMSE and almost perfect McFadden pseudo R-squared. This suggests the model is very well-fitted to the data.
- **Linear Regression** has a higher MSE and a moderate R-squared value, indicating that it does not fit the data as well as the Logistic Regression model.

If we plot predicted vs actual comparing linear and logistic regression we obtain

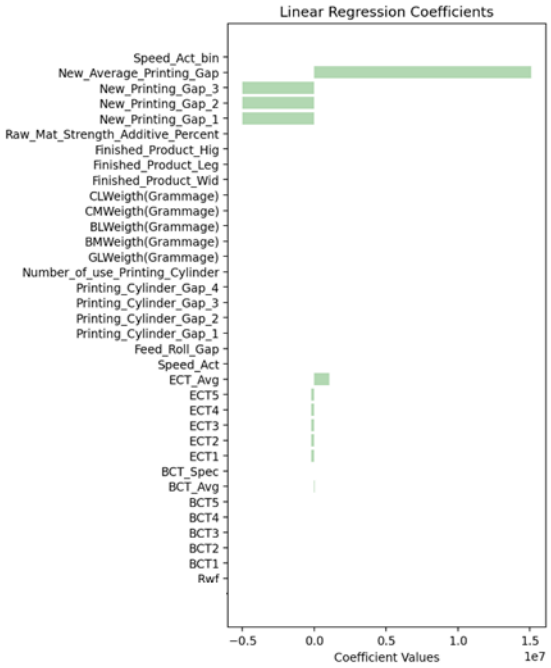

Linear Regression: Predicted vs Actual
Predicted Values (rules)
Actual Values (rules)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0.0
0.2
0.4
0.6
0.8
1.0


Logistic Regression - Actual vs Predicted with Logistic Line
Actual
Predicted
Logistic Line
Class Label (y_test and y_pred)
Feature Value (First Feature of X_test)
1.0
0.8
0.6
0.4
0.2
0.0
100
200
300
400
500
600

This is a confusion matrix, commonly used to evaluate the performance of a classification model by using logistic regression.



- There is only 1 false positive, meaning the model very rarely predicts "1" incorrectly when it should predict "0" and there are no false negatives, meaning the model does not miss any true positives.



Logistic regression appears to involve a broader range of variables than linear regression, which likely contributes to its higher accuracy. However, it's important to note that the dominant features influencing the Logistic Regression model are related to BCT (Box Compression Test) values, while the Linear Regression model is primarily driven by features related to machine printing cylinder gaps. This distinction suggests that each model relies on different aspects of the data to make predictions, with Logistic Regression capturing more complex relationships across multiple variables.

Conclusion

Corrugated box manufacturing can be easily disrupted if the final product does not pass the strength test. The consequence of producing weak boxes include financial losses due to shutting down the manufacturing line and a hold up in the supply chain among other things. To mitigate the chances of this instance, this research group applied various data mining techniques to create a model to predict box strength. The result of our efforts was the creation of a model that can accurately predict if a box is structurally sound regardless of the number of printing cylinders used.

Additional Links

https://github.com/vrahulrvce/Box_model

- This Github link contains the code we used to complete the final report