1. Introduction
   1. Injection Moulding Process

* Injection moulding is the most commonly used manufacturing process for the fabrication of plastic parts. A wide variety of products are manufactured using injection moulding, which vary greatly in their size, complexity, and application.
* The injection moulding process requires the use of an injection moulding machine, raw plastic material, and a mould.
* The plastic is melted in the injection moulding machine and then injected into the mould, where it cools and solidifies into the final part.
  1. Problem Statement
* How can data analytics be utilized for early detection and quality control in production processes to minimize rejections and improve efficiency through automated monitoring and control systems?
  1. Objectives
* Develop a data analytics framework that can effectively analyse real-time process data to detect and identify unacceptable variables or anomalies during production.
* Implement an automated system that can promptly respond to the detected issues, triggering necessary actions such as automatic rejection of the affected piece, thereby minimizing rejections and improving overall production efficiency.

1. Data Understanding and Preprocessing
   1. Understanding the data

The provided dataset contains information related to injection molding processes. The dataset includes the following attributes:

* Condition: The condition or type of the injection molding process.
* Cylinder heating zone 1-5: Temperature measurements of different heating zones of the cylinder.
* Maximum injection pressure: The maximum pressure applied during the injection molding process.
* Mould temperature control unit 1: Temperature measurement of the mold temperature control unit.
* Cycle time: The time taken to complete one cycle of the injection molding process.
* Injection time: The duration of the injection phase during the molding process.
* Dosage time: The time taken to dose the material for injection.
* Switch-over volume: The volume of material switched during the injection process.
* Material cushion: The material cushion thickness during injection molding.

Before analyzing the data, the following steps will be performed for data understanding and preprocessing.

* 1. Data Preprocessing

In the data preprocessing stage, various techniques are employed to clean and prepare the dataset for further analysis. The following techniques are used:

2.2.1 Handling Missing Data:

Missing data can affect the accuracy and reliability of the analysis. To address this issue, the following steps are performed:

1. Reading the Dataset: The dataset is loaded using the **pd.read\_csv()** function to create a DataFrame named **df**. The **info()** function is then used to provide an overview of the dataset, including the number of non-null values in each column.
2. Identifying Missing Values: The **isnull().sum()** function is used to determine the number of missing values in each column of the DataFrame. This helps in understanding the extent of missing data.
3. Filling Missing Values with KNN Imputation: The K-nearest neighbors (KNN) imputation technique is utilized to fill in the missing values. The **KNNImputer** class from the **sklearn.impute** module is used with **n\_neighbors=3** to impute missing values based on the values of the nearest neighbors. The 'Condition' column is stored separately, and the 'Condition' column is dropped from the DataFrame for imputation. After imputation, the DataFrame is converted back to its original form, and the 'Condition' column is inserted back.
4. Checking for Remaining Missing Values: The **isnull().sum()** function is used again to verify if any missing values remain in the DataFrame after imputation.

2.2.2 Handling Outliers:

Outliers can significantly impact statistical analysis and modeling. To handle outliers, the following steps are performed:

1. Removing Outliers with Median Replacement: The DataFrame **dff** is created by dropping the 'Condition' column. Numeric columns are selected using **select\_dtypes(include=np.number)**. For each numeric column, the Z-score method is employed to identify outliers. Any value with a Z-score above a threshold (e.g., 3) is considered an outlier. Outliers are replaced with the median value of the respective column.
2. Encoding the Condition Column: The 'Condition' column is encoded using label encoding to convert categorical values into numerical labels. The **LabelEncoder()** class from the **sklearn.preprocessing** module is used to perform this encoding. The encoded labels are inserted at the beginning of the DataFrame **dff**.

Finally, the preprocessed DataFrame **dff** is printed to observe the effects of missing data imputation, outlier replacement, and label encoding.

* 1. Discussion
* Data preprocessing is a crucial step in the data analysis process as it ensures the quality and reliability of the dataset. In this section, we discuss the key techniques applied during the data preprocessing stage and their significance in preparing the dataset for further analysis.
* Handling Missing Data: Missing data can arise due to various reasons such as measurement errors, data corruption, or system failures. Ignoring missing values can lead to biased results and inaccurate analysis. Therefore, the missing data handling techniques employed in this analysis are important.
* In this case, the dataset was examined for missing values, and it was found that certain columns had missing data. To address this, the KNN imputation technique was used. By utilizing the **KNNImputer** algorithm, missing values were imputed based on the values of the nearest neighbors. This approach helps to retain the structural integrity of the data and provides more accurate insights. The 'Condition' column was preserved separately, as it represents the categorical target variable, and after imputation, it was inserted back into the DataFrame.
* Handling Outliers: Outliers are data points that significantly deviate from the typical pattern of the dataset. These anomalous observations can skew statistical analysis and modeling results. Therefore, identifying and handling outliers is essential to ensure robust analysis.
* In this analysis, outliers were addressed using a median replacement approach. Z-score was calculated for each numeric column, and any value with a Z-score greater than a specified threshold (e.g., 3) was considered an outlier. The outliers were replaced with the median value of the respective column. This method helps in preserving the central tendency of the data and reducing the influence of extreme values.
* Label Encoding: Label encoding is employed to transform categorical variables into numerical labels. It allows algorithms to interpret and analyze categorical data effectively. In this case, the 'Condition' column, representing the type or condition of the injection molding process, was encoded using the **LabelEncoder()** function. This transformation facilitates the utilization of the categorical variable in subsequent analysis and modeling tasks.
* By employing these data preprocessing techniques, the dataset has been cleaned, missing values have been imputed, outliers have been addressed, and categorical variables have been encoded. The resulting preprocessed dataset, stored in the DataFrame **dff**, is now ready for further analysis and modeling tasks.

1. Statistical Analysis
   1. Mean/Median/range/std

In the statistical analysis section, the data is examined for different conditions: 'Normal1', 'Condition1', and 'Condition5'. The purpose of this analysis is to gain insights into the data and understand the distribution and variability of the variables under each condition. The following techniques are used:

**Descriptive Statistics**

Descriptive statistics provide summary measures that describe the central tendency, dispersion, and shape of the data. The **describe()** function is utilized to calculate various statistical measures, including the mean, median, standard deviation, minimum, 25th percentile, 75th percentile, and maximum. These measures help in understanding the average values, spread, and range of the variables.

**Range Calculation**

In addition to the standard statistical measures, the range is calculated separately for each variable. The range represents the difference between the maximum and minimum values of a variable. By including the range, we can gain insights into the overall span or variation of the data.

The statistical analysis is performed for each condition individually. This allows for a comparison of the statistical measures and ranges across different conditions. By examining the descriptive statistics, patterns and differences between conditions can be identified, helping to understand the impact of different conditions on the variables.

The use of statistical analysis techniques in this context is crucial for understanding the characteristics of the data, identifying potential outliers or anomalies, and gaining insights into the behavior of variables under different conditions. These insights can guide further analysis and decision-making processes related to data analytics and early detection through data analysis.

* 1. Discussion

The statistical analysis was performed on the dataset for three different conditions: 'Normal1', 'Condition1', and 'Condition5'. Each condition represents a specific scenario or set of factors that may influence the variables being analyzed. The analysis provides valuable insights into the characteristics and behavior of the variables under different conditions. Here is a discussion of the findings for each condition:

**Condition: Normal1**

The variables under the 'Normal1' condition exhibit the following statistical properties:

* Mean: The average value for each variable is around the following range:
  + Cylinder heating zone 1: 281.27
  + Cylinder heating zone 4: 310.303
  + Maximum injection pressure: 1537.27
  + Mould temperature control unit 1: 89.7854
  + Cycle time: 29.1022
  + Cylinder heating zone 5: 315.328
  + Injection time: 0.2346
  + Cylinder heating zone 2: 293.586
  + Dosage time: 2.96208
  + Cylinder heating zone 3: 308.944
  + Switch-over volume: 1.41387
  + Material cushion: 0.945836
* Median: The median value (50th percentile) for each variable is close to the mean, indicating a relatively symmetric distribution.
* Standard Deviation: The variables have moderate levels of dispersion around the mean, as indicated by the standard deviation.
* Range: The range provides insights into the overall variability of the variables, with values ranging from a minimum to a maximum. For example, the range for 'Cylinder heating zone 1' is 27.37.

**Condition: Condition1**

Under the 'Condition1' scenario, the variables show the following statistical characteristics:

* Mean: The mean values for each variable under 'Condition1' are as follows:
  + Cylinder heating zone 1: 289.436
  + Cylinder heating zone 4: 319.142
  + Maximum injection pressure: 1472.84
  + Mould temperature control unit 1: 89.8648
  + Cycle time: 28.7332
  + Cylinder heating zone 5: 326.413
  + Injection time: 0.237457
  + Cylinder heating zone 2: 304.411
  + Dosage time: 2.90353
  + Cylinder heating zone 3: 317.831
  + Switch-over volume: 1.39136
  + Material cushion: 0.920162
* Median: The median values for the variables are similar to their respective means, indicating a relatively symmetric distribution.
* Standard Deviation: The variables exhibit moderate levels of dispersion around the mean, as indicated by the standard deviation.
* Range: The range represents the overall variation in the variables, with values ranging from a minimum to a maximum. For example, the range for 'Cylinder heating zone 1' is 26.6.

**Condition: Condition5**

The statistical properties of the variables under 'Condition5' are as follows:

* Mean: The mean values for each variable under 'Condition5' are as follows:
  + Cylinder heating zone 1: 282.739
  + Cylinder heating zone 4: 308.395
  + Maximum injection pressure: 1565.07
  + Mould temperature control unit 1: 100.272
  + Cycle time: 28.5264
  + Cylinder heating zone 5: 314.873
  + Injection time: 0.232725
  + Cylinder heating zone 2: 295.403
  + Dosage time: 2.93798
  + Cylinder heating zone 3: 313.225
  + Switch-over volume: 1.40795
  + Material cushion: 0.968991
* Median: The median values for the variables are close to their respective means, indicating a relatively symmetric distribution.
* Standard Deviation: The variables display moderate levels of dispersion around the mean, as indicated by the standard deviation.
* Range: The range represents the overall variability of the variables, with values ranging from a minimum to a maximum. For example, the range for 'Cylinder heating zone 1' is 27.13.

By analyzing the statistical measures for each condition, we can observe differences and similarities in the behavior of the variables. These findings provide valuable insights into how the variables are affected by different conditions and can guide decision-making processes related to data analysis and quality control.

Please note that this discussion is based on the provided statistical analysis results. Further interpretations or conclusions may be drawn based on domain-specific knowledge and the context of the analysis.

## 4.Data Visualisation

## 4.1 Histogram and Line Chart

A picture containing text, screenshot, font, plot

Description automatically generated

A picture containing text, screenshot, font, plot

Description automatically generated

A graph with blue lines

Description automatically generated with low confidence

A picture containing text, screenshot, font, plot

Description automatically generated A picture containing text, screenshot, line, plot

Description automatically generated A picture containing text, screenshot, font, plot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, diagram, screenshot, plot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

4.2 Pair Plot

A picture containing text, handwriting, parallel, number

Description automatically generated

4.3 Discussion  
The 'Condition' variable in the dataset is influenced by various parameters. The correlation analysis using the **dff.corr()** function and Pearson's correlation coefficient reveals the following results:

1. Cylinder heating zone 1:
   * Correlation coefficient: -0.363
   * p-value: 8.90e-05
2. Cylinder heating zone 4:
   * Correlation coefficient: -0.320
   * p-value: 0.00063
3. Maximum injection pressure:
   * Correlation coefficient: 0.373
   * p-value: 5.43e-05
4. Mould temperature control unit 1:
   * Correlation coefficient: -0.151
   * p-value: 0.113
5. Cycle time:
   * Correlation coefficient: 0.230
   * p-value: 0.0152
6. Cylinder heating zone 5:
   * Correlation coefficient: -0.385
   * p-value: 2.95e-05
7. Injection time:
   * Correlation coefficient: -0.109
   * p-value: 0.253
8. Cylinder heating zone 2:
   * Correlation coefficient: -0.410
   * p-value: 7.92e-06
9. Dosage time:
   * Correlation coefficient: 0.250
   * p-value: 0.00816
10. Cylinder heating zone 3:
    * Correlation coefficient: -0.392
    * p-value: 2.12e-05
11. Switch-over volume:
    * Correlation coefficient: 0.202
    * p-value: 0.0331
12. Material cushion:
    * Correlation coefficient: 0.206
    * p-value: 0.0303

These results indicate the strength and direction of the relationships between each parameter and the 'Condition' variable. The correlation coefficient reflects the magnitude of the correlation, while the p-value helps assess the statistical significance of the relationship.

Based on the analysis, it can be concluded that the 'Condition' variable is influenced by parameters such as Cylinder heating zones (1, 4, 5), maximum injection pressure, cycle time, mould temperature control unit 1, injection time, dosage time, cylinder heating zone 2, cylinder heating zone 3, switch-over volume, and material cushion. The correlation coefficients and p-values provide insights into the degree of influence and statistical significance of these parameters on the 'Condition' variable.

A picture containing text, screenshot, pattern, square

Description automatically generated

## 5. Data Modeling and Error Calculation

### 5.1 Results and Observations

In this section, we present the results and observations obtained from the data modeling process. The data was divided into a training set and a test set using the train-test split technique, with a test size of 20%. Six machine learning models were applied to the training set, and their performance was evaluated using various metrics, including accuracy, precision, recall, and F1-score.

The following models were utilized for data modeling:

1. Logistic Regression:
   * Accuracy: 0.909
   * Precision: 0.929
   * Recall: 0.906
   * F1-score: 0.910
2. Decision Tree Classifier:
   * Accuracy: 0.802
   * Precision: 0.816
   * Recall: 0.802
   * F1-score: 0.794
3. Support Vector Classifier (SVC):
   * Accuracy: 0.828
   * Precision: 0.851
   * Recall: 0.809
   * F1-score: 0.822
4. K-Nearest Neighbors Classifier:
   * Accuracy: 0.819
   * Precision: 0.824
   * Recall: 0.802
   * F1-score: 0.805
5. Gaussian Naive Bayes:
   * Accuracy: 0.891
   * Precision: 0.908
   * Recall: 0.884
   * F1-score: 0.890
6. Multi-Layer Perceptron Classifier:
   * Accuracy: 0.874
   * Precision: 0.896
   * Recall: 0.874
   * F1-score: 0.875

Among these models, the best-performing one was Logistic Regression, which achieved an accuracy of 0.909, a precision of 0.929, a recall of 0.906, and an F1-score of 0.910. These results demonstrate the effectiveness of Logistic Regression in classifying the data.

### 5.2 Verification of Trained Model

To further optimize the performance of the selected models, a grid search technique was employed to tune their hyperparameters. Two specific models, Logistic Regression and Multi-Layer Perceptron Classifier, underwent hyperparameter tuning.

#### **Logistic Regression Hyperparameter Tuning**

For the Logistic Regression model, the following hyperparameters were tuned: penalty, C, and solver. The grid search was performed using a 5-fold cross-validation approach, and the evaluation metric used was accuracy. After the grid search, the best parameters for Logistic Regression were determined as follows:

* Penalty: l1
* C: 1.0
* Solver: liblinear

The best accuracy score achieved with these parameters was 0.928.

#### **Multi-Layer Perceptron Classifier Hyperparameter Tuning**

For the Multi-Layer Perceptron Classifier, the following hyperparameters were tuned: hidden\_layer\_sizes, activation, solver, and learning\_rate. Similar to the Logistic Regression model, the grid search was performed with a 5-fold cross-validation approach and accuracy as the evaluation metric. The best parameters for the Multi-Layer Perceptron Classifier were found to be:

* Hidden Layer Sizes: (100,)
* Activation: relu
* Solver: adam
* Learning Rate: constant

The best accuracy score achieved with these parameters was 0.892.

The hyperparameter tuning process aimed to optimize the models' performance by identifying the best combination of hyperparameters. These optimized models, with their tuned hyperparameters, can be considered well-calibrated and capable of achieving high performance on the given data.

6.Conclusion

In this analysis, we explored a dataset containing various parameters related to a specific condition. We performed data preprocessing, explored the dataset, and conducted several machine learning experiments to predict the condition based on the available features.

We started by loading the dataset and examining its structure, which included checking for missing values and ensuring the data types were appropriate for analysis. We also performed exploratory data analysis to gain insights into the distribution and relationships between the variables.

Next, we split the dataset into training and testing sets and applied feature scaling to normalize the features. We then trained several machine learning models, including Logistic Regression, Decision Tree Classifier, Support Vector Classifier (SVC), K-Nearest Neighbors Classifier, Gaussian Naive Bayes, and Multi-Layer Perceptron Classifier. We evaluated each model's performance using metrics such as accuracy, precision, recall, and F1-score.

Based on the evaluation results, the Logistic Regression model emerged as the best-performing model, achieving an accuracy of 0.909 and outperforming other models in terms of precision, recall, and F1-score. It exhibited the highest predictive capability for the given condition.

Furthermore, we examined the correlation between the 'Condition' variable and other parameters in the dataset. Using the Pearson correlation coefficient, we identified parameters such as cylinder heating zones, maximum injection pressure, cycle time, mould temperature control unit, injection time, dosage time, switch-over volume, and material cushion as influential factors affecting the 'Condition'. These parameters exhibited varying degrees of correlation and statistical significance.

Overall, this analysis provides valuable insights into understanding the relationship between the 'Condition' variable and the available parameters. The Logistic Regression model can serve as a reliable tool for predicting the condition based on the provided features. The identified influential parameters can guide further investigation and help optimize processes to improve the condition based on the associated factors.