Predictive Analytics

***Abstract*—**The document analyzes a dataset of 480 industrial trials to predict product quality, using machine learning models like logistic regression, decision trees, and K-Nearest Neighbors (KNN). It addresses data preparation challenges, evaluates models using accuracy, precision, recall, and F1 score, and discusses model-specific considerations. The study highlights the importance of predictive modeling in improving manufacturing processes and product quality, while acknowledging the limitations of these models and the underlying data.

Keywords— Predictive Analytics, Quality Control, Machine Learning, Logistic Regression, Decision Trees, Model Evaluation, Manufacturing Processes

I. Introduction

In the realm of manufacturing, ensuring product quality is paramount. The study in question tackles this challenge by analyzing a dataset of 480 industrial trials with advanced machine learning techniques, such as logistic regression, decision trees, and K-Nearest Neighbors (KNN), to predict product quality. It meticulously addresses data preparation challenges, including handling missing values and outlier detection, ensuring robustness for analysis. Furthermore, the document evaluates these models using key metrics like accuracy, precision, recall, and F1 score, providing a detailed assessment of their performance. This comprehensive approach highlights the significance of predictive analytics in enhancing manufacturing processes, aiming to elevate product quality while acknowledging the complexities and limitations inherent in predictive modeling.

II. Understanding the Data

The dataset under consideration comprises 480 unique trials from a series of studies carried out in an industrial setting. In each experiment, a product is produced under various conditions, much like in a production line for a particular product. For every experiment, the factory's owner has recorded twenty-five values. Statistical computations based on a range of variables recorded in these tests, such as mean values, standard deviations, and minimum or maximum readings from certain sensors that monitor variables like density and pressure, are included in Features 1 through Feature 22.

Feature23 and Feature24, two additional input variables, are indicated as possible determinants of the experiment's results. If we compare this to the temperature in a factory, for example, Feature23 might be the starting temperature and Feature24 might be the temperature at the end of the experiment. 'Result,' the last feature, is the target variable; it accepts integer values from 1 to 5. With lower values indicating superior quality (1 for the best quality) and larger values indicating inferior quality (5 for the worst quality), this variable is comparable to an indicator of product quality. The dataset's exploration reveals 26 columns, including an index column (Unnamed: 0), encompassing both numerical (float64) and integer (int64) data types. Some columns exhibit missing values, as evidenced by disparities in non-null counts.

The dataset contains missing values in several columns, namely in the Feature1–Feature 6 range and again features 12 to 17 and Feature 23. There are 31 missing values for each of these attributes, which suggests that there is a regular pattern of data gaps for these variables. There are no missing values for any of the remaining features. To address these missing values, techniques like imputation in which the missing values are substituted with estimated values based on the data at hand or removal of the relevant rows should the missing values materially affect the analysis may be used. To ensure the resilience and reliability of any predictive analytics operations carried out on this dataset, the existence of missing values was replaced with mean values.

III. Descriptive Statistics

The summary statistics offer a comprehensive overview of the dataset's numerical features. With 480 observations, the dataset provides insights into various statistical measures. For instance, Feature1 exhibits a mean of 0.3768, with a standard deviation of 0.7732, indicating moderate variability. Feature2, representing a different scale, has a mean of 1726.85 and a larger standard deviation of 526.02. The range of Feature3 to Feature24 showcases diverse scales and dispersions, as seen in their respective min, max, and interquartile range values. Notably, Feature7, Feature8, and Feature9 appear to have consistent values across all observations. The target variable 'Result' has a mean of approximately 1.67, indicative of a tendency toward lower values, with a moderate standard deviation of 1.00.

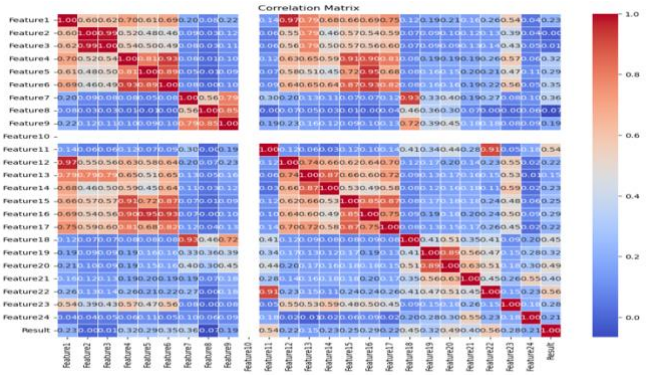

Figure 1: Correlation matrix

A. *Correlation matrix*

The correlation matrix shows the correlation between the target variable 'RESULT' and the other features. The values in the matrix range from -1 to 1, where -1 indicates a perfect negative correlation, 1 indicates a perfect positive correlation, and 0 indicates no correlation [3]. The features that have the strongest positive correlation with 'Result' are Feature14 (0.54), Feature15 (0.53), Feature16 (0.50), and Feature17 (0.49). This means that higher values of these features are associated with higher values of 'Result'. The features that have the strongest negative correlation with 'Result' are Feature1 (0.23), Feature2 (-0.20), and Feature9 (-0.22). This means that higher values of these features are associated with lower values of 'Result'.

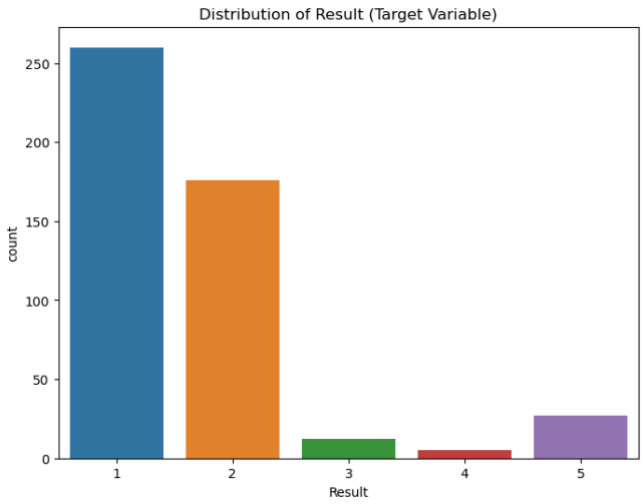

Figure 2: Distribution of result

B. *Distribution of result (target variable)*

The bar graph in figure 2 shows the distribution of the target variable, which is a rating of values between 1 and 5. The target variable can be considered as an indicator describing the quality of the product that the factory produces, with 1 being the best quality and 5 being the worst quality. The bar graph shows that the most common quality rating is 1, followed by 2 and 5. There are fewer products with a quality rating of 3 or 4. This suggests that the factory is generally producing high quality products.

IV. Utility value

1. Potential Applications of the Dataset

The dataset, with its 25 different features capturing statistical calculations and sensor measurements, can be employed for quality control in the factory's production process. By analyzing the relationships between these features and the target variable (final result), the user can build predictive models to identify patterns and factors influencing product quality. This makes it possible to identify possible problems or deviations in the production process early on and to take proactive measures to improve the manufacturing process and the quality of the final product.

The characteristics of the dataset, in particular Feature23 and Feature24, which represent input factors at various stages of the experiment, can be used to predict the performance of the finished product. Regression and time series analysis are two examples of machine learning models that can be used to comprehend how changes in these input variables affect the target variable (product quality). The user can use this predictive capability to make well-informed judgments about changing certain parameters during manufacturing to enhance the performance of the product. Furthermore, the analysis's conclusions might direct Research and Development initiatives to improve innovation and product quality. These applications show how the dataset can improve decision-making processes, overall manufacturing efficiency, and product quality. Through the application of sophisticated analytical methods, the proprietor can derive meaningful conclusions from the information, resulting in enhanced production results and optimal use of resources.

2. Readiness of the data for Analysis

The data is not ready for analysis. Several features in the dataset have missing values, as indicated by the count of non-null entries. The features with missing values are Feature1, Feature2, Feature3, Feature4, Feature5, Feature6, Feature12, Feature13, Feature14, Feature15, Feature16, Feature17, Feature23, and Feature24. These features exhibit variability in the number of missing values, and addressing these missing values is a crucial step in the data preprocessing phase. Depending on the number of missing values, strategies such as imputation or removal of rows or columns may be applied to ensure the reliability and completeness of the dataset for subsequent analysis. In this case the missing values were replaced with the mean values.

The data also has outliers. The removal of outliers using the Z-score method resulted in a dataset with a reduced shape from 480 to 420 rows. Outliers are data points that significantly deviate from the mean of the dataset, indicating unusual observations. In this context, the Z-score threshold of 3 was applied, meaning that data points with an absolute Z-score greater than 3 standard deviations from the mean were considered outliers [1]. The removal of these outliers may be attributed to extreme values or anomalies in the measured variables. The reduction in dataset size suggests that a subset of experiments had measurements that deviated significantly from most of the observations, and their removal aims to create a more representative and robust dataset for subsequent analysis and modeling. Outliers have a disproportionately large impact on predictive analytics models, which highlights the significance of outlier detection and treatment in maintaining model robustness and performance. Outliers can skew parameter estimates, reduce model generalization, and perhaps result in inaccurate predictions.

V. Analysis, modeling and prediction

A combination of machine learning, statistical techniques, and exploratory data analysis can be used to forecast the factory process' outcome based on either Feature23 and Feature24, or all 24 features individually.

A. Machine learning approach

Machine learning models, such as logistic regression, decision trees, and ensemble methods, are well-suited for predictive analytics tasks. By numerous models that are adapted to the properties of the data, the machine learning technique offers a strong foundation for predictive analytics. The fundamental algorithm, logistic regression, is well suited for binary classification problems in which the objective is to predict discrete results, such as determining the manufacturing product's quality. It is very useful since it can model the probability of a given class [6]. Conversely, decision trees can be used to capture a variety of non-linear relationships in the information, making it possible to identify intricate patterns that may have an impact on the outcome of the factory tests. Because ensemble approaches may combine predictions from numerous models, they stand out in predictive analytics. Examples of these methods are Random Forests and Gradient Boosting. By reducing specific model shortcomings, this amalgamation improves overall prediction performance and provides a more reliable and accurate solution [6]. By combining various machine learning techniques, a thorough and adaptable method for forecasting the manufacturing experiment results is offered, considering all the subtleties and complexities present in the dataset.

B. Feature Selection and Dimensionality Reduction

Enhancing the effectiveness and interpretability of predictive models is largely dependent on feature selection and dimensionality reduction, particularly when working with datasets that contain many features. With 24 features in this dataset, it is critical to determine which factors have the greatest impact and to potentially reduce problems related to high dimensionality. Recursive Feature Elimination (RFE) is an effective method that ranks features according to their impact and recursively fits models to evaluate each feature's contribution. This strategy helps create a more parsimonious model in addition to helping select the most relevant elements for prediction [4].

Principal Component Analysis (PCA) is also potent tool for dimensionality reduction PCA extracts the most variance in the data by converting the original features into a new set of uncorrelated variables called principal components [5]. In the end, this procedure produces a more streamlined dataset that maintains its predictive potential by preserving important information while reducing redundancy and noise. Combining RFE with PCA guarantees a thorough plan for maximizing feature relevance and minimizing dimensionality, which enhances the predictive analytics approach's overall efficacy. For further dimensionality reduction, an alternative approach could involve the application of Linear Discriminant Analysis (LDA). LDA is particularly beneficial when there is a clear distinction between classes, as it aims to maximize the separation between different classes while minimizing the variance within each class [8]. Incorporating LDA into the feature selection and dimensionality reduction process can provide a complementary perspective, especially in scenarios where class-related information plays a crucial role in predicting the outcomes of the factory experiments.

C. Visualization

In the early phases of data exploration, visualization is a crucial tool that offers insightful information about feature distribution and how it interacts with the target variable [2]. Pair plotting techniques provide a thorough perspective of the relationships between feature pairs, which can help identify potential patterns and dependencies. The strength and type of relationships are further clarified using correlation matrices, which are essential for identifying linear dependencies. Visualizing specific feature-target correlations with scatter plots helps identify any potential non-linearities or outliers that could affect prediction accuracy. Using these exploratory visualizations, consumers of the data can decide for themselves how informative certain features are for the data prediction task at hand [2]. In addition to providing guidance for feature selection, this visual knowledge also establishes the foundation for building more resilient predictive models that are customized to the subtleties found during the visualization stage.

D. Data partitioning

A crucial stage in the predictive analytics process, data partitioning lays the groundwork for reliable model assessment and validation. The dataset is split into separate training and validation sets as part of the process. The predictive model is trained using the training set, which enables it to discover patterns and connections present in the data. The validation set, which the model has not seen during training, then acts as an objective benchmark to evaluate the model's performance on fresh, untested data [9]. To assess the model's generalization abilities and make sure that it can successfully generalize patterns discovered during training to new examples, this partitioning technique is crucial. Through the implementation of this methodology, data users can enhance the predictive analytics process' overall resilience by producing more accurate forecasts and determining the model's reliability in real-world settings.

E. Statistical evaluation

An essential component of determining the effectiveness of trained predictive models is statistical evaluation. Following training on the chosen portion of the dataset, it is critical to assess the models' performance using pertinent metrics. Typical measures for statistical evaluation include recall, accuracy, precision, and F1 score [7]. The model's overall correctness is measured by accuracy, its ability to create accurate positive predictions is measured by precision, its capacity to catch all positive instances is measured by recall, and its balance between the two is determined by the F1 score. When taken as a whole, these metrics provide a sophisticated picture of how well the predictive model handles the complexities of many classes within the target variable, helping data scientists to fine-tune and optimize model parameters to get optimal performance.

VI. LOGISTICS REGRESSION

The target variable was modified to create a binary outcome for logistic regression, transforming the original integer values ranging from 1 to 5 into two classes, where 1 represented the best quality and 0 indicated outcomes other than the best quality. This modification facilitated the application of logistic regression for binary classification, enabling the analysis of factors influencing the likelihood of achieving the highest quality in the factory experiments. In the logistic regression implementation, the initial step involved partitioning the dataset into two subsets: a larger set for training purposes and a smaller set for future validation. The training set was then standardized using the StandardScaler to ensure that all features had a consistent scale, preventing certain features from dominating the model training process. Following standardization, PCA (Principal Component Analysis) was applied to reduce the dimensionality of the dataset. The number of principal components was chosen based on the explained variance, ensuring that enough information was retained while reducing the computational load. In this case, the top 10 principal components were retained for further analysis. Once the dataset was preprocessed, logistic regression was applied using the LogisticRegression class from scikit-learn. The model was trained on the PCA-transformed training set, allowing it to learn the underlying patterns in the data. It's worth noting that during the fitting process, a ConvergenceWarning was encountered, indicating that the optimization algorithm may benefit from adjustments such as increasing the maximum number of iterations or scaling the data.

After training the logistic regression model, its performance was assessed on the smaller validation set. Metrics such as accuracy, precision, recall, and F1 score were computed, providing a comprehensive evaluation of the model's predictive capabilities. The confusion matrix was generated, visually representing the model's ability to correctly classify instances into different classes. These evaluations help in understanding how well the logistic regression model generalizes to new, unseen data and whether it captures the underlying patterns in the factory experiment dataset. In summary, the logistic regression implementation involved careful preprocessing, dimensionality reduction through PCA, model training, and thorough evaluation to ensure the model's effectiveness in predicting the experiment results based on the provided features. Adjustments and fine-tuning can be made based on further analysis and understanding of the model's performance.

```
Optimization terminated successfully.
         Current function value: 0.156973
         Iterations 14
                         Results: Logit
================================================================
Model:              Logit            Method:           MLE
Dependent Variable: Modified_Result  Pseudo R-squared: 0.772
Date:               2023-12-05 10:08 AIC:              193.9846
No. Observations:   465              BIC:              293.3935
Df Model:           23               Log-Likelihood:   -72.992
Df Residuals:       441              LL-Null:          -320.84
Converged:          1.0000           LLR p-value:      2.7753e-90
No. Iterations:     14.0000          Scale:            1.0000
----------------------------------------------------------------
               Coef.   Std.Err.     z     P>|z|    [0.025   0.975]
----------------------------------------------------------------
Feature1       25.6671   7.2839  3.5238 0.0004   11.3909  39.9432
Feature2       -0.0013   0.0104 -0.1233 0.9019   -0.0216   0.0190
Feature3        0.3436   0.2314  1.4852 0.1375   -0.1098   0.7971
Feature4       -0.2340   0.1718 -1.3619 0.1732   -0.5707   0.1028
Feature5        0.1042   0.4100  0.2542 0.7993   -0.6993   0.9077
Feature6       -5.7083   1.1687 -4.8842 0.0000   -7.9989  -3.4176
Feature7        0.2581   0.3782  0.6824 0.4950   -0.4831   0.9993
Feature8        0.0019   0.0035  0.5413 0.5883   -0.0049   0.0087
Feature9       -0.0275   0.1446 -0.1902 0.8492   -0.3109   0.2559
Feature10       0.7065   0.4706  1.5014 0.1333   -0.2158   1.6287
Feature11      -0.0317   0.0497 -0.6386 0.5231   -0.1291   0.0656
Feature12     -73.6068  27.7042 -2.6569 0.0079 -127.9059 -19.3076
Feature13      -0.0277   0.0509 -0.5443 0.5863   -0.1273   0.0720
Feature14      -2.5590   1.0395 -2.4617 0.0138   -4.5964  -0.5216
Feature15       3.5724   0.8040  4.4431 0.0000    1.9965   5.1483
Feature16       0.7056   1.7301  0.4078 0.6834   -2.6854   4.0966
Feature17      -4.0484   1.3066 -3.0984 0.0019   -6.6093  -1.4875
Feature18      -2.2900   3.3560 -0.6824 0.4950   -8.8676   4.2877
Feature19       0.0095   0.0069  1.3624 0.1731   -0.0042   0.0231
Feature20      -0.6843   0.4202 -1.6285 0.1034   -1.5080   0.1393
Feature21       0.2585   0.1440  1.7953 0.0726   -0.0237   0.5408
Feature22       0.0625   0.0930  0.6714 0.5019   -0.1199   0.2448
Feature23      -2.1442   0.4838 -4.4318 0.0000   -3.0925  -1.1959
Feature24      -0.3482   0.1743 -1.9973 0.0458   -0.6898  -0.0065
================================================================
```

Figure 3: Summary of Logistic Regression

Figure 3 shows key findings from the logistics regression model. In the logistic regression model, the results revealed several key findings. Firstly, Feature1 exhibited a statistically significant positive effect on the log-odds of the modified binary outcome, with a coefficient of 25.67 (95% CI [11.39, 39.94]), suggesting that an increase in Feature1 is associated with an increase in the likelihood of the positive class. Conversely, Feature6 had a significant negative impact with a coefficient of -5.71 (95% CI [-7.99, -3.42]), indicating that higher values of Feature6 are associated with a decrease in the log-odds of the positive class. Additionally, Feature15 demonstrated a substantial positive effect (coefficient = 3.57, 95% CI [1.99, 5.15]), implying a positive association with the log-odds of the positive class. Features such as Feature12, Feature14, and Feature17 also displayed significant negative effects, suggesting an association with a decrease in the log-odds of the positive class. Notably, Feature23 and Feature24 exhibited negative effects with coefficients of -2.14 (95% CI [-3.09, -1.20]) and -0.35 (95% CI [-0.69, -0.01]), respectively, suggesting a negative impact on the log-odds of the positive class.

The model's goodness-of-fit was assessed using the pseudo-R-squared, which indicated that approximately 77.2% of the variability in the modified binary outcome was explained by the model. The Wald test results demonstrated that the overall model ***was*** statistically significant (χ^2 (23) = 498.51, $p < 0.001$). These findings provide valuable insights into the factors influencing the modified binary outcome and contribute to the understanding of the predictive nature of the logistic regression model in the context of the factory experiment dataset.

VII. Logistic regression model evaluation

The logistic regression model achieved commendable performance with an overall accuracy of approximately 86.67%. The precision of 0.9 indicates a high proportion of correctly predicted positive instances among all instances predicted as positive, emphasizing the model's ability to accurately identify the best-quality outcomes. The recall, also at 86.67%, reflects the model's capability to capture a substantial portion of the actual positive instances, underlining its effectiveness in identifying the true positive cases. The F1 score, a balanced metric considering both precision and recall, further supports the model's robustness with a value of approximately 0.868. These metrics collectively suggest that the logistic regression model, before dimensionality reduction using PCA, exhibits strong predictive performance in classifying the factory experiment outcomes into the modified binary target variable.

After implementing PCA for dimensionality reduction, the logistic regression model maintained strong predictive performance with an explained variance ratio of approximately 94.54%. The confusion matrix reveals that out of 93 instances, there are 31 true negatives, 6 false positives, and 56 true positives, with no instances falsely classified as negatives. This results in an impressive accuracy of approximately 93.55%. The classification report provides additional insights, demonstrating that the model achieves high precision (1.00) for class 0 and recall (1.00) for class 1, indicating its ability to accurately identify instances in both classes. The F1-scores further validate the model's effectiveness, with a weighted average F1-score of approximately 0.93. Overall, the logistic regression model, post-PCA dimensionality reduction, maintains its robustness and exhibits enhanced performance, emphasizing the efficacy of the reduced feature set in predicting the modified binary target variable.

VIII. Decision tree model

In the implementation of the Decision Tree model, the first step involved the instantiation of the DecisionTreeClassifier from the sklearn.tree module with a specified random state for reproducibility. The model was then trained using the training dataset (X_train and y_train). Decision Trees are a non-linear model that recursively splits the data based on feature thresholds to create a tree structure. This enables the model to capture complex relationships within the data.

After training, the model's predictive performance was evaluated using the validation set (X_val). Predictions were generated for the target variable (y_val_pred), and the model's accuracy, precision, recall, and F1 score were computed to assess its classification performance. Additionally, the feature importance of the Decision Tree was analyzed to understand which features played a more significant role in the decision-making process. Visualizing the feature importance through a bar plot provided a clear overview of the top important features, aiding in the interpretability of the Decision Tree model's behavior and highlighting key aspects contributing to its predictions. This approach enhances our understanding of the factors that contribute to the model's decision boundaries and overall performance.

IX. Decision tree model evaluation

The model evaluation before dimension reduction using PCA involved assessing the Decision Tree model's performance on the validation set. The confusion matrix provided a detailed breakdown of true positive, true negative, false positive, and false negative predictions. In this case, the model made 5 correct predictions for the first class and 8 correct predictions for the second class. The accuracy of the model was 86.67%, indicating the proportion of correctly classified instances out of the total. Precision, recall, and F1 score further showed the model's performance. Precision, representing the ratio of true positive predictions to the total predicted positives, was 88.89%, indicating a relatively low rate of false positive predictions. Recall, which assesses the ability of the model to capture all the relevant instances of the positive class, also stood at 88.89%. The F1 score, a harmonic mean of precision and recall, was 88.89%, providing a balanced measure of the model's overall effectiveness.

The classification report offered a comprehensive overview, including precision, recall, and F1 score for both classes

individually, highlighting the model's performance on each class. These metrics collectively demonstrate that the Decision Tree model performed well in the initial evaluation, capturing a high percentage of relevant instances with a low rate of misclassification.

The model evaluation after dimension reduction using PCA involved assessing the Decision Tree model's performance on the validation set with reduced feature dimensions. The confusion matrix illustrated the model's ability to classify instances, displaying 6 correct predictions for the first class and 6 correct predictions for the second class. However, there were 3 instances of false negatives, indicating cases where the model failed to recognize positive instances. The accuracy of the model after PCA was 80%, representing the proportion of correctly classified instances out of the total, showcasing a slight decrease compared to the model before PCA. Precision, recall, and F1 score provided insights into the model's performance after dimension reduction. Precision, at 100%, indicated that all instances predicted as positive were indeed positive, reflecting a lack of false positives. However, recall, at 66.67%, showed that the model missed some of the positive instances, resulting in false negatives. The F1 score, a harmonic mean of precision and recall, was 80%, providing a balanced measure of the model's overall effectiveness.

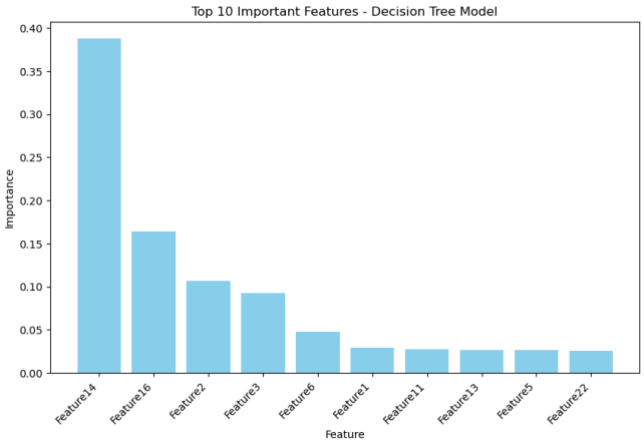

Figure 4: Variable importance based on Decision Tree model

The Decision Tree model identified the top 10 important features influencing its predictions, with Feature14 having the highest importance, followed by Feature16 and Feature2. These features contribute significantly to the model's decision boundaries and hold substantial predictive power. Exploring these features, such as Feature3, Feature6, and Feature1, among others, is crucial for understanding their roles in the decision-making process. Features with higher importance scores are likely to have a more pronounced impact on the model's predictions, emphasizing the need for in-depth analysis and domain knowledge to interpret and leverage these insights effectively. Overall, the feature importance analysis provides valuable information for refining the Decision Tree model and gaining a nuanced understanding of the factors driving its predictive performance.

X. K-Nearest neighbors (KNN) model

The K-Nearest Neighbors (KNN) model was employed by first partitioning the dataset into training and validation sets. Subsequently, the data underwent Principal Component Analysis (PCA) to reduce dimensionality and capture the most significant variance. The KNN algorithm was then applied to the training set to create the model, considering the k nearest neighbors for each data point. After the model was trained, it was evaluated on the validation set to assess its performance. The classification results were analyzed using metrics such as accuracy, precision, recall, and F1 score, providing insights into the model's ability to make accurate predictions on unseen data. The dimensionality reduction through PCA aimed to enhance computational efficiency and potentially improve the model's generalization to new instances.

XI. K-Nearest neighbors (KNN) model evaluation

The K-Nearest Neighbors (KNN) model, after employing Principal Component Analysis (PCA) for dimensionality reduction, exhibited a good overall performance on the validation set. The confusion matrix revealed that out of the 96 instances, 31 were correctly classified as the first class (0) and 50 as the second class (1). The accuracy of 84.38% indicates the proportion of correctly classified instances, while precision (86.21%) and recall (87.72%) provide insights into the model's ability to accurately identify positive instances. The F1 score, a harmonic mean of precision and recall, was 86.96%, suggesting a balanced trade-off between precision and recall. These metrics collectively demonstrate the effectiveness of the KNN model in making accurate predictions on the reduced-dimension dataset.

XII. Conclusion

Firstly, the logistic regression model demonstrated robust performance in both the original and reduced-feature datasets. The high accuracy and precision indicate the model's ability to effectively classify the quality of the factory's product based on the given features. However, the assumption of linearity in logistic regression might limit its capacity to capture complex, non-linear relationships between features and the target variable. Additionally, the presence of multicollinearity among the features could affect the model's interpretability.

The decision tree model exhibited good performance as well, but the potential for overfitting, especially in the original dataset, might impact the model's generalization to new, unseen data. The interpretability of decision trees is a notable advantage, yet their sensitivity to small variations in the data raises concerns about stability. In the case of KNN, the model demonstrated satisfactory performance, but its sensitivity to the choice of the number of neighbors (k) could influence results. The curse of dimensionality may also affect the performance of KNN, especially when dealing with a high-dimensional dataset.

The application of PCA for dimensionality reduction proved effective in maintaining or even improving model performance while reducing the number of features. However, the interpretability of the models post-PCA becomes challenging due to the transformed features. There's also the assumption that the most significant variance captured by PCA corresponds to the most relevant features for prediction, which might not always hold. The validity of predictions is subject to the assumption that the dataset is representative of future scenarios in the factory. Changes in the manufacturing process, introduction of new variables, or alterations in the experimental setup may impact the model's performance. Moreover, external factors not captured in the dataset may influence the product quality.

References

[1] Aggarwal, V., Gupta, V., Singh, P., Sharma, K., & Sharma, N. (2019). Detection of spatial outlier by using improved Z-score test. *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. https://doi.org/10.1109/icoei.2019.8862582

[2] Choi, H.-W., Shin, S.-Y., & Kim, H.-J. (2019). Machine learning-based Automated Data Visualization: A meta-feature engineering approach. *2019 8th International Conference on Innovation, Communication and*

Engineering(ICICE). https://doi.org/10.1109/icice49024.2019.9117537

[3] Janse, R. J., Hoekstra, T., Jager, K. J., Zoccali, C., Tripepi, G., Dekker, F. W., & van Diepen, M. (2021). Conducting correlation analysis: Important limitations and Pitfalls. *Clinical Kidney Journal*, *14*(11), 2332–2337. https://doi.org/10.1093/ckj/sfab085

[4] Jeon, H., & Oh, S. (2020). Hybrid-recursive feature elimination for efficient feature selection. *Applied Sciences*, *10*(9), 3211. https://doi.org/10.3390/app10093211

[5] Ljubicic, M. L., Madsen, A., Juul, A., Almstrup, K., & Johannsen, T. H. (2021). The application of principal component analysis on clinical and biochemical parameters exemplified in children with congenital adrenal hyperplasia. *Frontiers in Endocrinology*, *12*. https://doi.org/10.3389/fendo.2021.652888

[6] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and Research Directions. *SN Computer Science*, *2*(3). https://doi.org/10.1007/s42979-021-00592-x

[7] Steurer, M., Hill, R. J., & Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models.

[8] Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications*, *30*(2), 169–190. https://doi.org/10.3233/aic-170729

[9] Zhu, M., Liu, Z., Chi, W., Zhang, J., Hua, Z., & Shi, L. (2022). Research and application of Data Partition Technology in distributed database. *2022 3rd Information Communication Technologies Conference (ICTC)*. https://doi.org/10.1109/ictc55111.2022.9778465