**Predictive Analytics [PGR304]**

2023 Autumn

Candidate 2002

**Abstract**

**Introduction**

A quantitative evaluation by using metrics and visualization, also a qualitative discussion by incorporating interpretation of the results and critical reflection.

**Methodology**

**Understanding the data**

what kind of dataset is this? Provide some context by saying a little about what kind of object we are talking about, are they correlated or not…..

In the given dataset, the ‘Feature’ and ‘Result’ columns are variables, and rows are the factory experiment records. All the variables are numeric in the dataset.

Plotting the distribution of input variables, particularly the input variable of ‘Feature 23’ and ‘Feature24’ in the case, is important for several reasons:

Understand variation. It helps in understanding how these input variables vary across different experiments. For instance, if ‘Feature23’ represents the temperature at the start of the experiment, its distribution can display the range of temperatures experiments are being started at.

Identifying Anomalies. The distribution plots can help identify any anomalies or outliers in the input variables which might affect the quality of the product.

Relationship with Target. By plotting these features against the target variable, we can begin to see if there is any visible trend or pattern indicating how these input variables might be affecting the outcome.

For input variables like Feature 23 and 24, considering they might have a direct impact on the result, it could be beneficial to plot them to see if there are an obvious patterns or clusters which could be related to the quality of the product.

A collage of graphs

Description automatically generated

A screenshot of a video game

Description automatically generated

A screenshot of a video game

Description automatically generated

The heatmap visualization gives a quick identifying how each feature is related to the others.

Darker Red Square indicates a strong positive correlation. This means that as the value of one feature increases, the value of the other feature also increases.

Darker Blue Square suggests a strong negative correlation. As the value of one feature increases, the value of the other feature decreases.

Lighter Colors indicates weaker correlations.

Based on the Pearson correlation matrix, here are the top 10 features most highly correlated with the target variable 'Result':

Feature22: Correlation coefficient of 0.558414

Feature11: Correlation coefficient of 0.541232

Feature20: Correlation coefficient of 0.492456

Feature18: Correlation coefficient of 0.452594

Feature21: Correlation coefficient of 0.400463

Feature7: Correlation coefficient of 0.360749

Feature19: Correlation coefficient of 0.317561

Feature6: Correlation coefficient of 0.296198

Feature23: Correlation coefficient of 0.255241

Feature4: Correlation coefficient of 0.241024

For numerical features, histograms or density plots are common ways to visualize distributions. For the target

**Utility value**

Based on the understanding of the data, the dataset can be used for several applications, particularly because it contains experimental results from a production process with various measurements:

Quality Control and Improvement.

By predicting the quality of the product, the factory can proactively identify and correct issues in the production process, leading to a more consistent product quality. In the given dataset, the target variable in the ‘Result’ column can classify the product quality from 1 to 5.

Process Optimization. Analysis of the dataset could reveal which features (process parameters) are most influential on product quality, enabling process engineers to fine-tune the production line for optimal performance.

Predictive Maintenance. If the features include data from machine sensors, predictive models could help anticipate equipment failures before they occur by correlating sensor readings with experiment outcomes that had poor quality results.

Cost Reduction. By understanding which features and variables impact the quality, the factory can focus resources on monitoring and controlling critical factors, potentially reducing the cost associated with over-processing or excessive quality assurance testing.

Based on the initial exploration of the given dataset, the dataset is ready for analysis.

Missing values have been detected and replaced by median;

Descriptive statistics, feature distributions and feature classification have been examined by machine learning;

Feature correlations are printed in a matrix and plotted in a heatmap.

However, before proceeding with further analysis and modeling, the following pre-processing steps are considered:

Feature Scaling. Machine learning algorithms like KNN and SVM are sensitive to the scale of the data. Standardization (z-score normalization) or Min-Max scaling might be required.

Dimensionality Reduction. There are 24 features in the given dataset, especially correlated ones. Techniques like PCA could be used to reduce the number of features while retaining most of the information.

Handling Class Imbalance. From the distribution of target variable ‘Result’, the variable is imbalanced. Hence, techniques like SMOTE for oversampling the minority class or stratified sampling during train-test split can be considered.

Feature Selection. Although we’ve looked at correlations, more sophisticated feature selection methods can be applied to retain only the most relevant features for the modeling task.

**Analysis, Modeling, Prediction**

Based on the given dataset contains 480 experiments with 24 features each, 5 classes classification in the result, also with net-numerical data at the size of 170kb size, it is a relatively small dataset. Traditional machine learning approaches are sufficient for the training, prediction and analysis.

Logit Model

KNN

SVM

Naive Bayes

Decision Tree

A black screen with many small colored squares

Description automatically generated

PCA

LDA

**Result and evaluation**

The evaluation of selected models

Models Evaluation

Accuracy

Precision

Recall

F1-Score

Confusion Matrix

**Discussion**

**Discovery**

1. Abbott. D., Komp. W.J. (2014). *Applied Predictive Analytics.* Chapter 3, Page 46-50 (Data distribution and skewness)
2. Abbott. D., Komp. W.J. (2014). *Applied Predictive Analytics.* Chapter 4, Page 113-114 (PCA)
3. Abbott. D., Komp. W.J. (2014). *Applied Predictive Analytics.* Chapter 7, Page 204-205 (ANOVA)
4. Abbott. D., Komp. W.J. (2014). *Applied Predictive Analytics.* Chapter 8, Page 214-229 (Decision Trees)
5. Abbott. D., Komp. W.J. (2014). *Applied Predictive Analytics.* Chapter 8, Page 254-263 (KNN)
6. Brownlee. J. (2020). *Multinomial Logistic Regression With Python.* [online] Available at: <https://machinelearningmastery.com/multinomial-logistic-regression-with-python/> (Logit Model)
7. Brownlee. J. (2020). *Linear Discriminant Analysis for Dimensionality Reduction in Python.* [online] Available at: <https://machinelearningmastery.com/linear-discriminant-analysis-for-dimensionality-reduction-in-python/> (LDA)
8. Geeksforgeeks. (). *How to create a seaborn correlation heatmap in Python?* [online] Available at: <https://www.geeksforgeeks.org/how-to-create-a-seaborn-correlation-heatmap-in-python/>
9. Turney.S. (2021). *Pearson Correlation Coefficient (r) | Guide & Examples.* [online] Available at: <https://www.scribbr.com/statistics/pearson-correlation-coefficient/>
10. Scikit Learn (2007-2023). *Linear and Quadratic Discriminant Analysis with covariance ellipsoid.* [online] Available at: <https://scikit-learn.org/stable/auto_examples/classification/plot_lda_qda.html> (LDA)
11. Scikit Learn (2007-2023). *Nearest Neighbors Classification.* [online] Available at: <https://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html>
12. Scikit Learn (2007-2023). *Decision Trees.*  [online] Available at: <https://scikit-learn.org/stable/modules/tree.html>