

Faculteit Industriële
Ingenieurswetenschappen

master in de industriële wetenschappen: elektronica-
ICT

Masterthesis

Optimizing the request for quotation process with machine learning-based connector clustering

Alexander Lemmens

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: elektronica-ICT

PROMOTOR :

dr. Nikolaos TSIOGKAS

PROMOTOR :

Mvr. Rosa ROCHA

Gezamenlijke opleiding UHasselt en KU Leuven



Universiteit Hasselt | Campus Diepenbeek | Faculteit Industriële Ingenieurswetenschappen | Agoralaan Gebouw H - Gebouw B | BE 3590 Diepenbeek

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE 3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE 3500 Hasselt



2023
2024

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: elektronica-
ICT

Masterthesis

Optimizing the request for quotation process with machine learning-based connector clustering

Alexander Lemmens

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: elektronica-ICT

PROMOTOR :

dr. Nikolaos TSIOGKAS

PROMOTOR :

Mvr. Rosa ROCHA



KU LEUVEN

Preface

This master's thesis marks not only the culmination of my academic journey in Electronics-ICT Engineering at UHasselt/KULeuven but also a profound personal and professional growth experience. Over the course of this research, I have navigated the complexities of the automotive industry, particularly focusing on the intricacies of the Request for Quotation (RfQ) process within the context of supply chain management.

The process of exploring machine learning techniques to streamline this aspect of the industry has been both challenging and exhilarating. It pushed me to apply theoretical knowledge, innovate, and think critically about real-world applications. The insights gained through this endeavor have not only shaped my academic perspective but also my aspirations for my future career.

I owe a debt of gratitude to many who have supported me throughout this journey:

- Prof. Dr. Ir. Nikolaos Tsiogkas, my thesis advisor, for his invaluable guidance, patience, and expertise. His keen insights and encouragement were crucial in navigating the challenging aspects of this research.
- Rosa Rocha for her expert advice and feedback, which were immensely helpful in refining my analysis and arguments.

This thesis would not have been possible without the contributions and support of each one of these individuals and groups.

Contents

Preface

List of figures

Abstract

Abstract in Dutch

1	Introduction	12
1.1	Context.....	12
1.1.1	Yazaki business overview.....	12
1.1.2	Global Process Management System (GPMS)	13
1.1.3	Request for Quotation (RfQ)	14
1.1.4	Wire harness connector	15
1.1.5	Data science in the industry	17
1.2	Problem statement.....	17
1.3	Objectives	18
1.4	Structure.....	18
2	Methodology	20
2.1	Data preprocessing	20
2.1.1	Data cleaning.....	20
2.1.2	Data reduction.....	22
2.1.3	Data transformation.....	24
2.2	Machine learning clustering.....	25
2.2.1	Unsupervised clustering.....	25
2.2.2	Supervised clustering	28
2.2.3	Combining clustering methods	30
2.2.4	Distance measurements	30
2.3	Validation and testing.....	30
2.3.1	Validation techniques.....	30
2.3.2	Testing.....	33
3	Results and discussion	35
3.1	Visualizing the data before using machine learning.....	35
3.2	Performing K-means clustering on the 2D and 3D plot.....	36
3.3	Unsupervised connector clustering	39
3.3.1	K-means.....	39
3.3.2	DBSCAN.....	42
3.3.3	GMM	43
3.3.4	Decision tree	45

3.4	Clustering with true labels and comparative analysis	47
3.4.1	Standardization	47
3.4.2	Adding weights to the characteristics	51
3.4.3	Different initialization techniques	53
3.4.4	Balanced K-means	55
3.4.5	Handling outliers.....	57
3.5	Most similar wire-harness connector	59
4	Conclusion and feature work	58
	References	60

List of Figures

Figure 1.1 Wire harness in a motor vehicle	12
Figure 1.2 Yazaki business overview.....	13
Figure 1.3 Overview of the RfQ process in the automotive industry	14
Figure 1.4 Drawing of a standard wire harness connector	15
Figure 2.1: K-means clustering of connectors based on the total poles and weight.....	21
Figure 2.2: Visual representation of the PCA method.....	23
Figure 2.3: Example of one-hot encoding of colors.....	25
Figure 2.4: Visual overview of the DBSCAN method.....	28
Figure 2.5: Example of a decision tree.....	29
Figure 2.6: Confusion matrix of K-means results.....	31
Figure 2.7: (a) confusion matrix before adjustment and (b) confusion matrix after adjustment.....	32
Figure 3.1: 2D plot of connectors based on total poles and weight.....	35
Figure 3.2: 3D plot of connectors based on terminal size, total poles, and weight.....	36
Figure 3.3: Silhouette score graph.....	37
Figure 3.4: K-means clustering of connectors based on total poles and weight.....	37
Figure 3.5: Silhouette score graph.....	38
Figure 3.6: 3D plot of connectors after using K-means.....	39
Figure 3.7: Silhouette score graph.....	40
Figure 3.8: 2D PCA plot of K-means clustering.....	41
Figure 3.9: 2D t-SNE plot of K-means clustering.....	41
Figure 3.10: t-SNE plot for three groups.....	42
Figure 3.11: PCA plot of the DBSCAN results of 9 clusters.....	43
Figure 3.12: PCA plot of GMM results of 17 clusters.....	44
Figure 3.13: t-SNE plot of GMM results of three clusters.....	44
Figure 3.14: 3D plot of K-means clusters.....	45
Figure 3.15: Decision tree based on the K-means results.....	46

Figure 3.16: K-means clustering with standardization.....	48
Figure 3.17: Confusion matrix with standardization.....	49
Figure 3.18: K-means clustering without standardization.....	50
Figure 3.19: Confusion matrix without standardization.....	50
Figure 3.20: K-means clustering with weighted characteristics.....	52
Figure 3.21: Confusion matrix with weighted characteristics.....	52
Figure 3.22: K-means with custom initial centroids.....	54
Figure 3.23: Confusion matrix with custom initial centroids.....	54
Figure 3.24: Balanced K-means.....	56
Figure 3.25: Confusion matrix of balanced K-means.....	57
Figure 3.26: K-means clustering after handling outliers.....	58
Figure 3.27: Confusion matrix after handling outliers.....	59
Figure 3.28: K-distance graph.....	60
Figure 3.29: Second round of clustering using DBSCAN.....	61
Figure 3.30: Third round of clustering using DBSCAN.....	62

Abstract

In the automotive industry, it is imperative for suppliers to swiftly respond to a Request for Quotation (RfQ) from an Original Equipment Manufacturer (OEM) with accurate pricing to sustain a competitive advantage. However, the current approaches for the RfQ process are considerably time-consuming. This master's thesis explores the application of machine learning techniques to identify the most similar connectors within wire harnesses, thereby optimizing the RfQ procedure.

The proposed approach included four stages. The initial stage was data preprocessing, which involved data cleaning, transformation, and reduction to prepare the feature vectors for the clustering techniques' inputs. Subsequently, the connectors were grouped into 17 clusters using K-means clustering and the Gaussian Mixture Model (GMM). In the third stage, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was performed to categorize similar connectors within the same group. The final step then included an iterative execution of the DBSCAN to find the most similar connectors depending on the number of connectors needed.

A dataset consisting of connectors with verified labels validates the proposed approach. The results indicate an overall 62% accuracy rate among 1,000 connectors. Consequently, the effectiveness of the approach has been demonstrated. The thesis also includes an analysis of the performance of different techniques and discussions on future improvements.

Abstract in Dutch

In de auto-industrie is het noodzakelijk voor leveranciers om snel te reageren op een Request for Quotation (RfQ) van een Original Equipment Manufacturer (OEM) met accurate prijzen om een concurrentievoordeel te behouden. De huidige benaderingen voor het RfQ-proces zijn echter tijdrovend. Deze masterproef onderzoekt de toepassing van machine-learningtechnieken om de meest vergelijkbare connectoren binnen kabelbomen te identificeren en zo de RfQ-procedure te optimaliseren.

De voorgestelde aanpak bestond uit vier fasen. In de eerste fase werden de gegevens voorbereid, waarbij de gegevens werden opgeschoond, getransformeerd en verkleind om de feature vectors voor te bereiden op de inputs van de clusteringstechnieken. Vervolgens werden de connectoren gegroepeerd in 17 clusters met behulp van K-means clustering en het Gaussian Mixture Model (GMM). In de derde fase werd Density-Based Spatial Clustering of Applications with Noise (DBSCAN) uitgevoerd om vergelijkbare connectoren binnen dezelfde groep te categoriseren. De laatste stap omvatte vervolgens een iteratieve uitvoering van DBSCAN om de meest vergelijkbare connectoren te vinden, afhankelijk van het aantal benodigde connectoren.

Een dataset bestaande uit connectoren met geverifieerde labels valideert de voorgestelde aanpak. De resultaten wijzen op een nauwkeurigheid van 62% onder 1000 connectoren. De effectiviteit van de aanpak is dus aangetoond. De thesis bevat ook een analyse van de prestaties van verschillende technieken en discussies over toekomstige verbeteringen.

Chapter 1

Introduction

1.1 Context

1.1.1 Yazaki business overview

Yazaki is a global automotive supplier of wire harnesses. An automotive requirement of an Original Equipment Manufacturer (OEM) like Toyota or BMW dictates the entire production process in the automotive supplier sector. An OEM can require particular parts, like wire harnesses. A wire harness in the automotive industry assembles different components, such as cables, terminals, connectors, etc., which are distributed in the vehicle, and each has a function [1], [2]. A wire harness is shown in Figure 1.1.

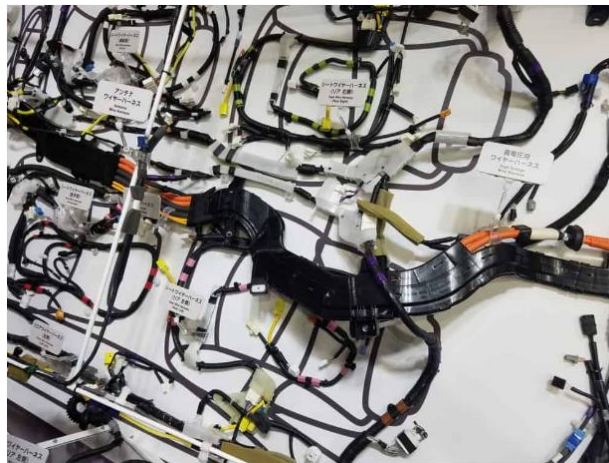


Figure 1.1: Wire harness in a motor vehicle [2]

When a supplier company like Yazaki receives such a requirement, the supplier's corporate management develops a plan to meet the request while including costs, technology, market dynamics, etc [1].

Corporate management will adjust the business process by using the feedback from different departments within the company. The feedback includes performance, progress, and potential risks [1].

Ultimately, the supplier aims to deliver products that meet the OEM’s requirements. After completing the project, the components have been mass-produced and sent to the OEM to be assembled into the vehicle [1]. Figure 1.2 shows the Yazaki business overview.

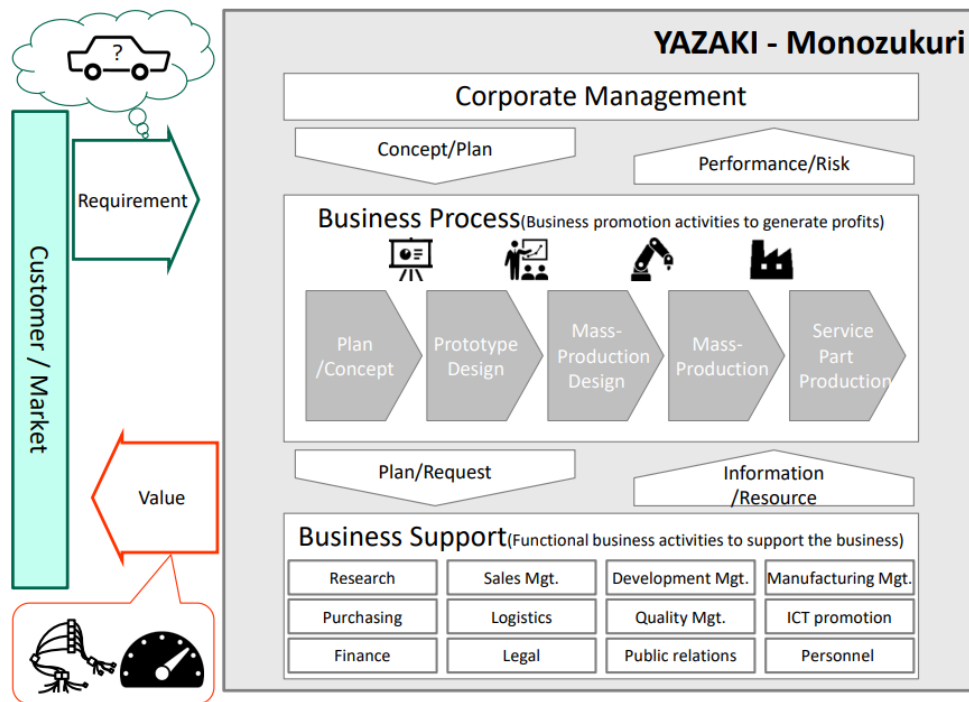


Figure 1.2: Yazaki business overview [1]

The wire harnesses and, thus, electrical connections make it possible for the vehicle to work. As time progresses, the wires that connect sensors, actuators, indicators, etc. become more complex. The wire harnesses in a car are the third most expensive components after the engine and chassis. Recent studies estimated that by 2030, about half the cost of the car will come from the electric system, whereas in 2010, it was 30%. The manufacturing and design of the harnesses are complex processes that need specialized software, a significant amount of human labor, and some automation [3].

1.1.2 Global Process Management System (GPMS)

To increase company efficiency, promote the achievement of Yazaki policy, and meet the demands of society and customers, the GPMS was established.

The values of GPMS are based on the demands of society and customers, such as high quality, low cost, proper delivery time, and being environmentally friendly. GPMS was introduced because there were a few problems with previous processes, namely confusing and conflicting rules of multiple processes, so improvement was needed.

This system establishes a global Performance Management Process (PMS) structure aligned with customer-specific requirements. Secondly, it comprises seven phases (phases 0-6), from

planning to mass production. Also, individual work processes are executed, and performance is assessed in each phase. It can be concluded that GPMS ensures project results and processes that determine the performance [4].

1.1.3 Request for Quotation (RfQ)

One of the most important phases of the GPMS is the RfQ process. In the automotive industry, this process serves as the foundation for establishing a business relationship between an OEM and a supplier. Creating a new part or component made by an automotive supplier begins with an RfQ, published by the OEM. The supplier then has to react to it with a response. This response consists of a technical and commercial offer. The technical part consists of the product requirements, and the commercial part consists of the price, annual volumes, cost breakdown, etc [5].

In addition to the objective of an RfQ to create a business relationship, it has a substantial financial influence on suppliers, determining their profitability and operational dynamics for years. Recent studies uncovered that certain areas within the RfQ processes have some difficulties, and better RfQ processes are needed because automotive components are evolving quickly, and supplier networks are becoming more complex. For example, extensive specifications and performance orders are necessary for important components in wire harness manufacture to achieve accurate pricing and component selections. An example of such an important component is a connector [6].

Furthermore, time restrictions are complicating the RfQ response process. It is possible that suppliers have to handle multiple RfQs at once. This burdens their resources and could affect the quality of their reaction to each request [6].

Because of this high-pressure setting, it is necessary to manage data functionally and make smart use of available information for the supplier to speed up the response. Recent studies indicate that there is hope thanks to the development of software solutions like data science, AI, and machine learning for handling the RfQ response process [6].

Figure 1.3 shows an overview of the RfQ process in the automotive industry and displays the different steps that need to be taken in various departments to complete the RfQ process. It also presents the average time needed to go from data preparation to export generation.

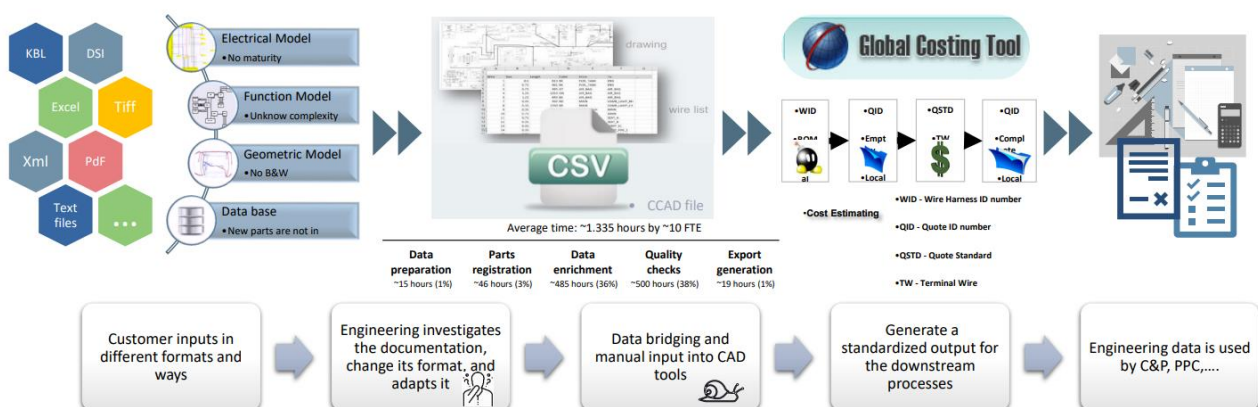


Figure 1.3: Overview of the RfQ process in the automotive industry [7]

1.1.4 Wire harness connector

A connector is an important component in the wire harness business, and clear specifications are necessary for accurate pricing and component selection for the response to the RfQ.

Definition and role

Automotive electrical connectors have the task of connecting various parts of a vehicle's wire harness. Besides facilitating electrical connections, they also control data transmission and signal integrity. In addition, these connectors are essential to the efficient operation of vehicle systems, which improves the car's overall security, performance, and reliability [8].

The connectors must meet the conditions imposed to cope with the conditions in the vehicles. The automotive environment demands connector performance, and in some cases, the connectors must be able to withstand changes in temperature, humidity, corrosion, etc [9]. Figure 1.4 illustrates a drawing of a standard connector.

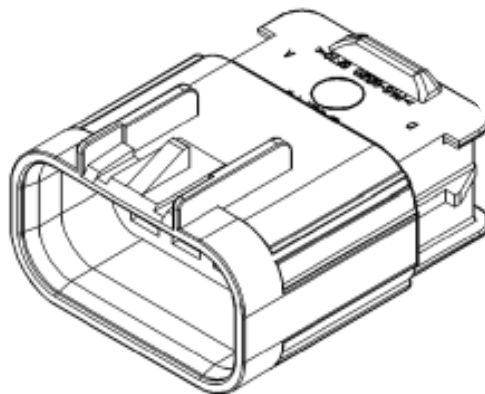


Figure 1.4: Drawing of a standard wire harness connector

Importance in the wire harness business

Regarding reliability, connectors are regularly a weak point in electrical and electronic systems located in vehicles. An example of a common failure is high or intermittent contact resistance faults. A major cause of these failures is degradation [9].

On a technical level, connectors are important within the RfQ process. When an OEM sends an RfQ to an automotive supplier, the supplier must look at the connectors that the OEM needs. The suppliers then see if they have or make the components themselves or if they should purchase them. This process for connectors is not easy because of the many important attributes that must be met to ensure the proper operation of a connector within the wire harness. Since response time plays an important role for the RfQ in the automotive business, it could be profitable to automate this [10].

Financially, connectors are also an essential component of the wire harness industry due to the high margin that can be made. Additionally, wire harness manufacturers need price and costing

specifications to make precise cost estimations. An automated costing tool would remove any misquotes or mistakes, helping manufacturers to increase profitability [10].

Characteristics of connectors

Connectors have a large number of attributes, and because time is an important factor during the RfQ process in the wire harness business, only the characteristics needed to know if a particular connector can be used that the OEM requested.

Some of the key connector attributes in the RfQ process are cavity family, total poles, weight, and gender.

Firstly, each cavity family has a specific type of terminal pins or contacts that can be matched with the cavities. It should be a match so the connector is compatible with the specific wire and pin type [11].

Another critical attribute is the poles of a connector. The number of poles refers to the amount of electrical connections, or in other words, each wire belongs to one pole. The number can vary depending on the location or functionality of the specific connector. To ensure the stable operation of a wire harness, it is necessary to have a good insertion of the wires into the poles so that a good electrical connection is maintained in difficult situations [12].

Today, car manufacturers and consumers want vehicles that consume less fuel and thus emit less. So, by reducing the vehicle's weight, emissions can be reduced. Because of this, the weight of connectors plays an important role. This characteristic depends on other attributes, such as cavity family and the number of poles, which will be essential for classifying connectors based on their attributes [13].

A final essential characteristic discussed here is the gender of a wire harness connector. This attribute refers to the type of electrical interface it provides. They can be divided into male and female types. For example, male connectors, which are equipped with pins, are matched to fit into female types. The female types contain sockets. This way, connecting correctly ensures connections and prevents electrical failures [14].

Complexity and variability

The wire harnesses are one of the bulkiest parts of the car. One of the main reasons why they are so complex is the need for reliable connections within a vehicle that must not lose their function even in extreme situations such as strong vibrations, high-temperature changes, and humidity. Connectors also provide indirect safety by connecting various sensors and actuators that provide vehicle safety [14].

Because of physical constraints, connectors need to accommodate different functional demands. Research suggests a transformation is necessary within the functionality of connectors so that they can handle multiple functions in limited space. An example of a connector with multifunctionality is a hybrid connector interface. This connector proves both a signal and a power connection [14].

1.1.5 Data science in the industry

Data-driven decision-making

Research shows that process optimization and innovation are possible within the automotive industry through the use of data science and data-driven decision-making. These optimizations and innovations can increase productivity and profit by making better decisions with fewer mistakes [15].

For example, data-driven decision-making can be used in the logistics department within the automotive industry. This technique helps to improve procurement tactics and supply chain logistics, which results in time and cost savings [15].

Another example is in the sales and marketing departments, where automotive firms can improve their sales and marketing strategy by getting a better view of market trends and what customers prefer through data-driven insights [15].

Challenges and considerations

The automotive industry's obstacles in data science must be overcome to integrate them into the company's operations [16].

One potential difficulty is data-driven culture. Creating a culture is an obstacle, and a lot of training is needed to change traditional business paradigms towards data-centric decision-making models [32].

Another obstacle is data governance. Effective data governance and a good data strategy are necessary to utilize data science fully. To apply this, there is a need for procedures and guidelines for managing the data and integrating it into the corporate strategy vision [16].

1.2 Problem statement

As mentioned earlier, the RfQ processes must be improved because of rapidly evolving components and increasingly complex supplier networks. The wire harness components must be accurately priced and selected. Furthermore, time restrictions are becoming more important during the RfQ response process. During this process, a lot of data is processed manually. In addition, there exist inaccuracies because of assumptions about parts and data processes and the high complexity of interfaces. Moreover, the preparation of the RfQs faces long lead times, and the process disrupts the day-to-day schedule and, therefore, causes delays in current business [6].

The connector is one of the most important parts of a wire harness regarding the RfQ process. Clear specifications are necessary for accurate pricing and component selection for the response to the RfQ. Connectors are most important because of their reliability and the high margin that can be made. Additionally, connectors contain a total of 123 different characteristics. Because of this, it takes a long time for the engineers to find interchangeable connectors as fast as possible for the RfQ response.

1.3 Objectives

The main objective of this thesis is to find similar connectors based on their characteristics to tackle the problems regarding the RfQ process for connectors using machine learning-based clustering methods.

However, before using clustering techniques, this thesis aims to create preprocessed datasets ready to be processed using clustering methods. Different clustering techniques will be applied to the preprocessed dataset and compared to investigate which ones are best for grouping the connectors based on their characteristics without knowing the type of connectors in the dataset.

Another objective is to suggest a clustering approach using the K-means method specifically to group wire harness connectors while having additional information about the connector types to validate and compare various approaches.

Finally, this study aims to find the advantages and limitations of the suggested approaches and discusses feature work to improve these suggestions.

1.4 Structure

Chapter 2 explains the methodology used to achieve the objectives of this master's thesis to find solutions to the different stated problems. The methodology chapter is separated into data preprocessing steps, machine learning clustering methods, and validation and testing.

Chapter 3 aims to guide the reader in understanding the results and discussion of this study. It contains three main parts. First, it performs three different clustering methods to group the wire harness connectors on a dataset containing approximately 6,500 without any true labels. Second, five comparisons were made to find the best possible approach for grouping the connectors using the K-means algorithm on a dataset containing 1,020 connectors, including true labels. Third, the results contained findings on the most similar connectors based on their characteristics.

The conclusion and feature work follow, including the objective, the key findings, implications, limitations, and suggestions for future research. The last pages contain the references used in this thesis.

Chapter 2

Methodology

2.1 Data preprocessing

Data preprocessing is a step needed before using machine learning models on the data if the raw data contains quality issues such as missing values, outliers, features with varying scales, etc. If these issues are not resolved, the results of the clustering algorithms will be suboptimal because of inefficient clustering, increased computation time, and misleading results. Data preprocessing involves cleaning, reducing, and transforming the data so the quality is enhanced and the clustering models performed on the data will be more meaningful and accurate. For this thesis, this process can be subdivided into three groups: data cleaning, reduction, and transformation [18].

2.1.1 Data cleaning

Data cleaning consists of handling outliers and missing values. Outliers and missing values are unwanted in the dataset for this project because they can lead to inaccurate analysis [18].

Handling outliers

An outlier in the dataset is considered a data point that lies an abnormal distance from other data points. The occurrence of outliers can skew the results [19].

An example of outliers in a dataset in this thesis is high-voltage connectors, a connector group with a significantly higher weight. Figure 2.1 below shows a graph of connectors after using k-means on the dataset. The data points in yellow are clustered as one group: high-voltage connectors. To protect these connectors from being deleted as outliers, they can be filtered before handling the outliers.

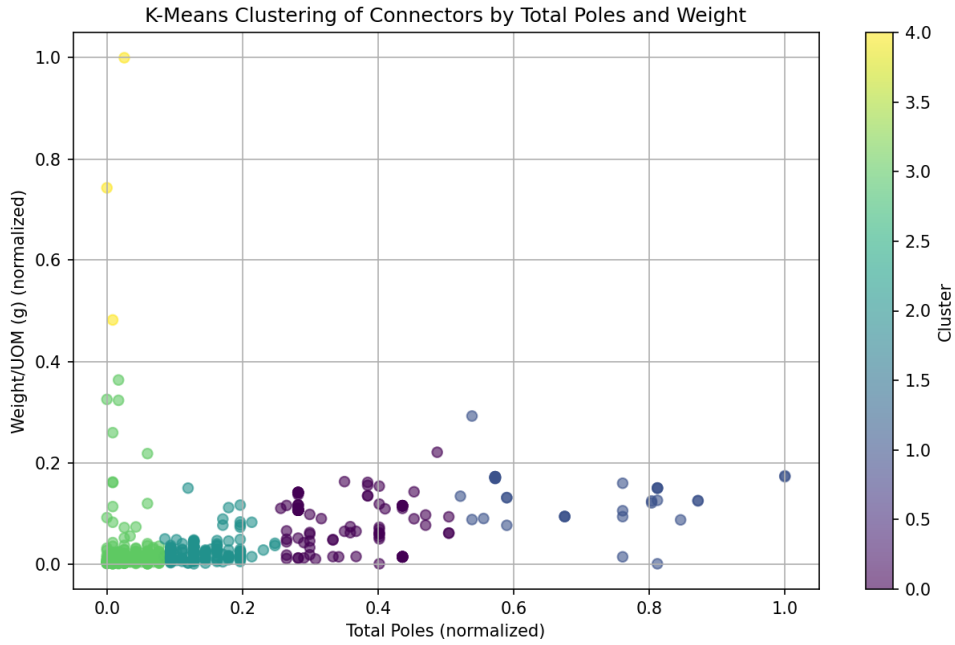


Figure 2.1: K-means clustering of connectors based on the total poles and weight

To avoid misleading conclusions, the method used for handling outliers was calculating the mean of all the numerical values per connector. Next, the mean of those mean values was determined to serve as a reference point. Finally, the relative position of each connector from the reference was determined by calculating the difference between the mean per connector and the reference point. The choice was made to remove the top 20% of connectors regarding the highest difference. The other 80% of the dataset was retained.

Handling missing values

Handling missing values in the dataset is necessary for this project because, for example, K-means clustering requires complete data to form clusters. The mean, mode, and median imputation methods have been used in this project to handle the missing values.

When to use missing value imputation techniques depends on the dataset size and the amount of missing data. For this project, a dataset is considered small if it contains less than 1,000 data points and large if it comprehends more than 10,000 data points. A small percentage of missing values is typically less than 10%. Although these numbers can vary depending on the application and different factors like dimensionality and the available computational power.

When the dataset is large, the impact of imputed values will be minimal. On the other hand, for small datasets, imputation can distort the data more noticeably. Additionally, if a small amount of data is missing, mean, mode and median imputation can be an effective solution. Conversely, using imputation methods when a large proportion of the data is missing can introduce inaccurate representation. In this project, the rows with one or more missing values were deleted when working with a large proportion of missing values or a small dataset.

The mean imputation method is used for numerical values. It calculates the mean of a column in the dataset and replaces the missing values with the calculated average value in that column. The

advantage is that it is easy to implement. Still, the resulting variance estimate for this feature can be underestimated when the number of missing values is significant for that variable [20].

The median imputation is similar to the mean method but calculates the median value instead of the average value. In the case of a column where the values are similar, this method is preferred over the mean values as it is more robust to outliers [20].

In the instance of categorical data, such as the color of a connector, mode imputation is most appropriate to apply. This method will substitute missing values with the most frequent value in that column.

2.1.2 Data reduction

Data reduction involves reducing the form of the dataset while maintaining its essential features. This step is vital since clustering algorithms such as k-means can be computationally intensive, and reducing the dataset size can improve resource-intensiveness [18]. Furthermore, plotting results after using a clustering method with more than three dimensions is impossible, which can be solved by dimensionality reduction techniques. The most suitable features were selected because not all connector characteristics are relevant for grouping the parts.

Feature selection

The choice of a connector influences the overall performance of an automotive wire harness and, thus, the security and performance of the vehicle [21].

A wire harness connector has 123 unique characteristics, and 15 of the most necessary connector attributes for grouping those wire harness components were selected. These 15 features were chosen based on their importance in finding the best matching connectors.

The 15 necessary connector attributes to replace a connector are the catalog type, cavity family, color, gender, hybrid poles, locking, material, layout, number of open poles, sealed, minimum and maximum temperature, terminal size, total poles, and weight.

They were selected based on the standard guide engineers in Yazaki use in choosing the ideal connector for an application when designing a wire harness, where one is the first and most important step and seven is the least important step in the guide:

1. the requirements for electrical parameters,
2. safety parameters,
3. mechanical parameters,
4. connection modes,
5. installation modes and appearance,
6. environmental parameters,
7. termination methods

Dimensionality reduction

Reducing the dimensionality of a dataset has the advantage of decreasing the number of variables, which makes it less time to compute the models and adds storage space. On top of that, redundant data will be removed, which will help the clustering algorithms to work more efficiently, and it will contribute to removing the noise from the dataset because it focuses on the most informative features [22].

Two dimensionality reduction techniques were implemented: Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). These two methods were utilized as visualization techniques to plot the clustering results. The disadvantage of these techniques for plotting the results is the complexity of interpretation because the transformed features are not as interpretable as the original features.

PCA

PCA is a statistical dimensionality reduction method that will map the data in a higher dimensional space into a lower dimensional space. At the same time, the variance after the transformation should be maximum. After reducing the dimensions, the goal is to maintain the most significant patterns or correlations between the variables [23].

The variables of the original dataset are linearly combined to create the orthogonal axes or principal components. They are arranged in decreasing order of importance, and their variance equals the original dataset's. The assumption is made that the information of the dataset is encoded in the variance of the features, which means the higher the feature's variance, the higher the information it holds [23]. Figure 2.2 illustrates the usage of PCA on the original data, represented in a 2D space.

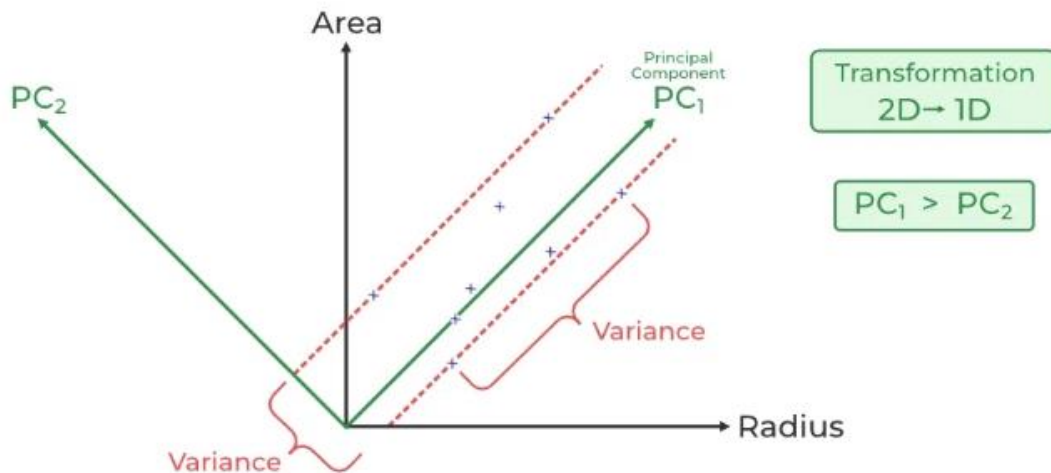


Figure 2.2: Visual representation of the PCA method [23]

The figure has a 2D space with two axes, showing the data points. PC_1 (Principal Component 1) is the direction where the variance is the majority. PC_2 is the second component and represents the direction of the second most variance in the data. The objective is to transform the 2D data into 1D, which will be transformed along PC_1 to preserve the variance as much as possible. The amount of variance in the figure is visualized by the red lines and brackets, which is at most for PC_1 [23].

The main disadvantage of utilizing PCA to reduce the number of dimensions is that it is limited to linear transformations and thus may not apprehend non-linear relationships in the data [24].

T-SNE

t-SNE is a nonlinear dimensionality reduction technique, unlike PCA, which is a linear technique. While PCA focuses on maximizing the variance after reducing the dimension, t-SNE preserves the relationships between the data points after the transformation to a lower-dimensional space. This makes t-SNE an effective approach for visualizing high-dimensional data, but it is more computationally expensive than PCA [24].

The similarity measures between pairs of data points are determined by t-SNE. Consequently, it optimizes two similarity measures.

Firstly, in the original state of the dataset, the pairwise similarity between all data points is determined using a Gaussian kernel. The more data points are close together, the higher the probability they will be put together. Thereafter, the t-SNE algorithm will map the high-dimensional data points of the original dataset onto a lower-dimensional space while keeping the pairwise similarities. It minimizes the sum of the differences in similarities using gradient descent until the lower-dimensional space is in a stable state [24].

2.1.3 Data transformation

Even after data cleaning and reduction, the adapted dataset may not be ready for further analysis. Data transformation will change the format or structure of the dataset so it becomes appropriate for clustering methods. In this thesis, data transformation involves possibly normalization, standardization, and encoding of categorical values [18].

Normalization

The numerical values were transformed to a fixed range between zero and 1 in this project using Min-Max scaling. This ensures that no single feature with a larger range of values dominates the results of the clustering methods. This ensures the same scale for each feature.

The formula for normalization using Min-Max scaling is shown below, where x' is the normalized value, x is the original value, \min is the lowest value of that feature, and \max is the highest numerical value of that feature [25].

$$x' = \frac{x - \min}{\max - \min} \quad (2.1)$$

Standardization

Standardization or Z-score normalization will transform the numerical data so that the feature has a mean of zero and a standard deviation of one. This method was not always used as a data transformation method because it is particularly useful when the data has a Gaussian distribution.

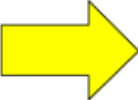
The formula for standardization is indicated below, where y' is the standardized value, y is the original value, μ represents the mean value, and σ is the standard deviation [25].

$$y' = \frac{y - \mu}{\sigma} \quad (2.2)$$

Encoding categorical values

In this thesis, before using clustering methods on the dataset, the categorical features were converted to binary vectors using the one-hot encoding method.

Figure 2.3 below illustrates an example of the one-hot encoding method [26].



Color	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow	0	0	1

Figure 2.3: Example of one-hot encoding of colors [26]

The feature color, for example, is a categorical variable, and a separate column per possible value for this feature is added. The presence of a value is then indicated by putting a one in the corresponding column and a zero in all the others, which generates a binary vector.

2.2 Machine learning clustering

Machine learning clustering methods will be used to group the connectors based on their characteristics in this project. There are multiple algorithms, each with its own advantages and disadvantages. It is a technique that aims to identify groups or clusters in which observations are more similar to each other than observations aligned to different clusters [27].

Clustering is useful in several situations, including data mining, document retrieval, image segmentation, and pattern classification. Nevertheless, there may be little prior information (e.g., statistical models) available about the data, and the decision-maker must make as few assumptions about the data as possible. Under these restrictions, clustering methodology is particularly appropriate for exploring interrelationships among the data points to assess their structure [28].

The clustering methods used in this thesis are K-means, balanced K-means, Gaussian Mixture Model (GMM), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and decision trees.

2.2.1 Unsupervised clustering

Unsupervised cluster analysis involves automatically discovering natural grouping in the data without knowing the groups of the data points in the dataset. This means the data used is

unlabeled and unclassified. In this study, the dataset contained around 10,000 connectors without prior knowledge of their type.

K-means clustering

K-means clustering, one of the most widely used clustering algorithms, will assign data points to one of the K clusters. The assignment is done based on the distance of a data point from the center of the clusters.

The K-means algorithm performs iterative calculations to optimize the positions of the centroids or centers of the clusters. These calculations are stopped when the values stop changing or the predefined number of iterations has been achieved. It works as follows:

1. Beforehand, the number of clusters or centroids K must be decided. After, it will randomly initialize those points. There are different methods for initializing these initial positions, and several will be discussed in this thesis.
2. Every data point will be categorized to its closest centroid using a distance measure method, such as the Euclidean distance.
3. After, the centroids' coordinates are updated.
4. The first three steps are repeated for the number of iterations or until the centroids no longer change significantly [29].

The limitation of this K-means method is that it is sensitive to the initialization of the clusters; because of this, the results of performing the same K-means algorithm on the same dataset can vary. Also, this clustering method requires a predefined number of clusters, which can be difficult to determine for unlabeled data.

Balanced K-means clustering

The balanced K-means clustering method requires all clusters to be the same size. This method was used to ensure the data points were equally distributed across all the clusters and was compared to the normal K-means method for the connector dataset.

Gaussian Mixture Model (GMM)

The GMM is a probabilistic model that tries to model the data as a combination of multiple Gaussian distributions [31]. This method determines the probability of each data point belonging to a certain cluster [32].

The main difference between K-means and GMM is that K-means is a hard clustering method, which means it groups each data point into one cluster. On the other hand, GMM is a soft clustering method, which means it checks the probability that a data point is associated with a specific cluster [27], [18].

The GMM algorithm works as follows:

1. First, the GMM's parameters are initialized. An option is to use the previous results obtained from the K-means method, which can provide an initial starting point.

2. Secondly, the expectation step is performed. It involves calculating the likelihood of each data point belonging to each component of the Gaussian distribution using the current estimates of the method parameters.
3. After the maximization step, the parameters, such as the means and mixture weights, are adjusted according to the calculations in step two. This will maximize the probability of the dataset.
4. The expectation and maximization steps are repeated until there is convergence in the likelihood value [32].

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN is a clustering method that groups data based on how close data points lie together depending on two parameters of the algorithm. The points that lie in low-density regions are classified as outliers. This means that not all points are assigned to a cluster [33], [34], [35].

The DBSCAN algorithm contains four steps:

1. It defines two parameters: 'eps' (epsilon, the maximum distance between two points to be considered neighbors) and 'minPts' (the minimum number of points required to form a dense region).
2. Second, the algorithm classifies the points into three kinds of points. Firstly, core points are those with the least 'minPts' neighbors within 'eps' distance. The second kind is border points, which are points within the 'eps' distance of a core point but have fewer than 'minPts' neighbors. Lastly, there are noise points, points that are neither core nor border points.
3. After classifying the points, all the core points are assigned to clusters, and the clusters are expanded by iteratively adding density-reachable points (points within 'eps' distance from any point in the cluster).
4. Finally, the algorithm is terminated when no more new points are added to any cluster.
5. The main differences between DBSCAN and the previously mentioned clustering techniques are that DBSCAN does not require the number of clusters because it determines the number of clusters based on the data density and can distinguish noise and clusters [33], [34], [35], [36].

Figure 2.4 shows a simple visual overview of the DBSCAN algorithm.

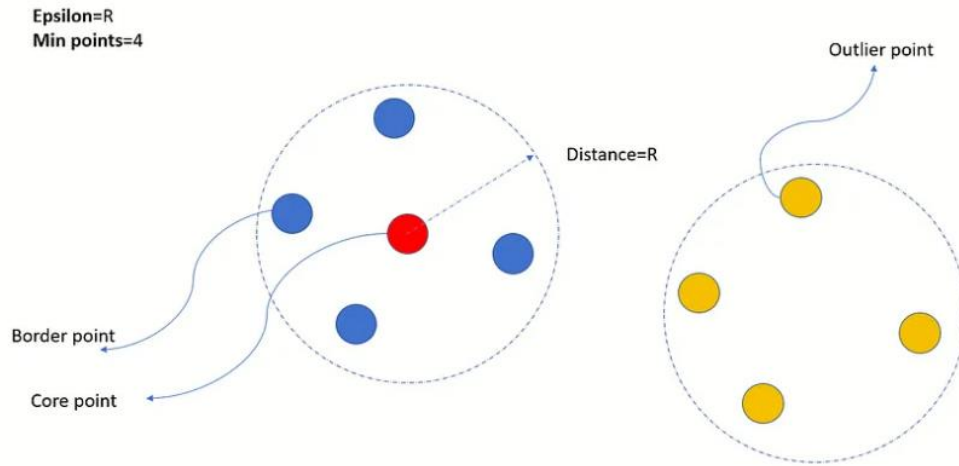


Figure 2.4: Visual overview of the DBSCAN method [35]

The red point, the core point, has at least a minimum of data points (Min points) within a specified radius (epsilon). The border points in blue are within the radius of the core point but do not have enough neighboring points to be a core point itself. The outlier data points, in yellow, do not lie within the predefined radius of the core point [35].

Initialization strategies

Because the K-means algorithm is very sensitive to the initialization of the clusters, three methods were used in this thesis to verify which one is best for the clustering of wire harness connectors.

The first method used for the initial placement of centroids is random initialization, which will select k numbers of data points as the initial cluster centers. However, this can lead to suboptimal solutions. To improve the random initialization method, the algorithm was run, for example, ten times with different initializations, and the best result was picked based on verification criteria such as the sum of squared distances. Although, with prior knowledge, the seeding method can be applied to make the initial guesses. Knowing the number of groups, one connector per group was used for the first centroid. The connector picked per group is a connector with the average of every feature, so it is ensured that the picked connector represents that group best [37].

2.2.2 Supervised clustering

The first phase of this master's thesis focused on unsupervised clustering to a dataset of approximately 6,500 connectors. This dataset didn't involve true labels or prior knowledge of their grouping. Multiple clustering techniques were employed to find the natural grouping of the connectors based on their characteristics.

Another dataset of approximately 1,000 connectors with true labels was used. This means this dataset included information on the type of connector. This data was used to compare and verify

the clustering techniques. In total, there are 17 types of connectors in Yazaki, namely: lamp, digital signal, infotainment, MOST, modular, joint, high power, squib, injections, door, standard, optical, PCB/header, safety, transmission, special, and RF connectors.

Based on the results of one of the clustering methods, decision trees were employed to make it explainable by identifying key features that differentiate clusters.

Decision trees

An effective method to interpret resulting clusters is the use of decision trees. This method analyses the clustering results and determines which features are most influential in defining each cluster. On top of that, it clearly represents the logic behind the clustering. It can be used to communicate clustering results to stakeholders. This method is suitable for applications where it is important to understand why some decisions were made in the clustering methods [38], [39].

Decision trees split the dataset into subsets based on the features that give the greatest information gain or impurity reduction (e.g., Gini impurity) [38], [39]. The Gini impurity is the probability of misclassifying an observation. An example decision tree that works this way is illustrated below:

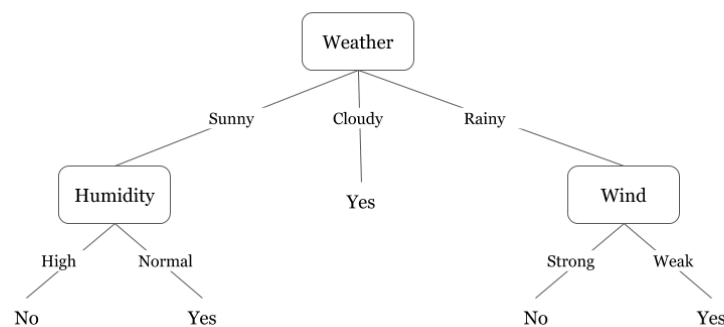


Figure 2.5: Example of a decision tree [39]

Figure 2.5 is a decision tree and starts at the root node. In this example, the root node is ‘weather,’ a feature resulting in the highest information gain. Second, the internal nodes, ‘humidity’ and ‘wind’ in this case, represent another feature based on a decision. The branches from each node represent the possible outcomes of the decisions. Lastly, the leaf nodes represent the final classification. Each path from the root to the leaf node represents a unique decision rule [39].

2.2.3 Combining clustering methods

Multiple clustering methods were combined to meet one of the objectives of this study, namely, to identify the most similar connectors based on their characteristics to find interchangeable connectors. The strengths per clustering technique can be leveraged this way.

For the first clustering round, the K-means technique was implemented. With prior knowledge of the connector types, the dataset was grouped into 17 clusters. Then, the data points of one of the groups were extracted, and a second round of clustering with DBSCAN was performed on that group. This second round found smaller and denser clusters within that group. DBSCAN sets the optimal number of clusters and assigns the labels the algorithm finds back to the connectors. DBSCAN was chosen because it finds the optimal amount of clusters and because it can identify noise points. The second round could be performed more times to find even smaller clusters.

2.2.4 Distance measurements

Distance measurements can be performed to quantify the similarity between connectors. In this study, the Euclidean distance metric is chosen. In a multi-dimensional space, this metric is a straight line between two points. To find the distance, it calculates the square root of the sum of the squared spaces between the coordinates of the two points [40]. Given (x_1, y_1) , the coordinates of the first point, and (x_2, y_2) , the coordinates of the second point, the Euclidean distance d is defined as:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (2.3)$$

This distance measurement can be used in the initial group assignment. When a new connector is added to the dataset, the Euclidean distance between the new connector and the center of each cluster can be calculated. The new connector belongs to the cluster whose center has the smallest distance to the connector. Also, this method can be used to assign the connector to a sub-cluster whose centroid is closest in terms of the Euclidean distance.

2.3 Validation and testing

Validation and testing are crucial to evaluate the different clustering techniques applied to the dataset of the wire harness connectors. This ensures the results meet the objectives and provide meaningful insights.

2.3.1 Validation techniques

The validation techniques used in this thesis are silhouette scores and confusion matrices.

The silhouette score is an internal validation. This type of validation is without true labels, and the internal metrics are used to evaluate the quality of the clustering based on the structure of the data itself. This score measures how similar an object is to its own cluster compared to other clusters. This means that the higher this score, the better defined the clusters are. The score ranges from -1 to 1. A score of one defines the clusters as well-separated, and the sample is far from its neighboring cluster. A score of zero suggests overlapping clusters, or the sample is on or very close to the decision boundary separating two neighboring clusters, and a score of -1 indicates highly overlapping clusters or incorrect clustering [41]. The calculation of the silhouette score s can be done by using the following formula:

$$s = \frac{p-q}{\max(p,q)} \quad (2.4)$$

where p is the mean distance to the points in the nearest cluster, and q is the average distance to other points in the same cluster. The mean of all the individual scores is taken to calculate the overall silhouette score. This overall score is a good measure of clustering quality [41].

The second validation technique performed in this thesis is confusion matrices. A confusion matrix helps visualize a clustering technique's performance compared to true labels. It illustrates the connections between the predicted cluster assignments and the actual types. For this matrix type, the diagonal values should be highest, as the diagonal values represent correct grouped data points. Off-diagonal elements highlight misclassifications. Figure 2.6 below shows a confusion matrix used in this master's thesis.

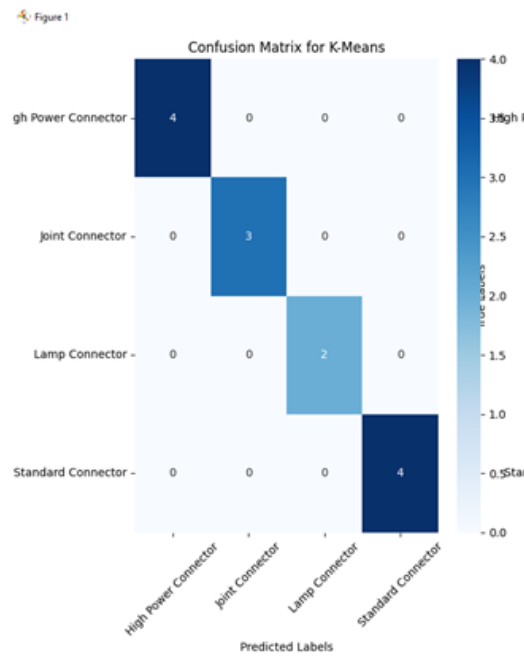


Figure 2.6: Confusion matrix of K-means results

Figure 2.6 illustrates the confusion matrix for the K-means clustering algorithm. The x-axis shows the labels predicted by the K-means algorithm. These categories list are the predicted connector types. The y-axis shows the true labels of the connectors. In this case, there are no misclassified connectors.

The Hungarian algorithm was performed to minimize the misclassification algorithm and ensure an optimal label assignment. The reason why this algorithm was used is shown by a problem represented in Figure 2.7 below.

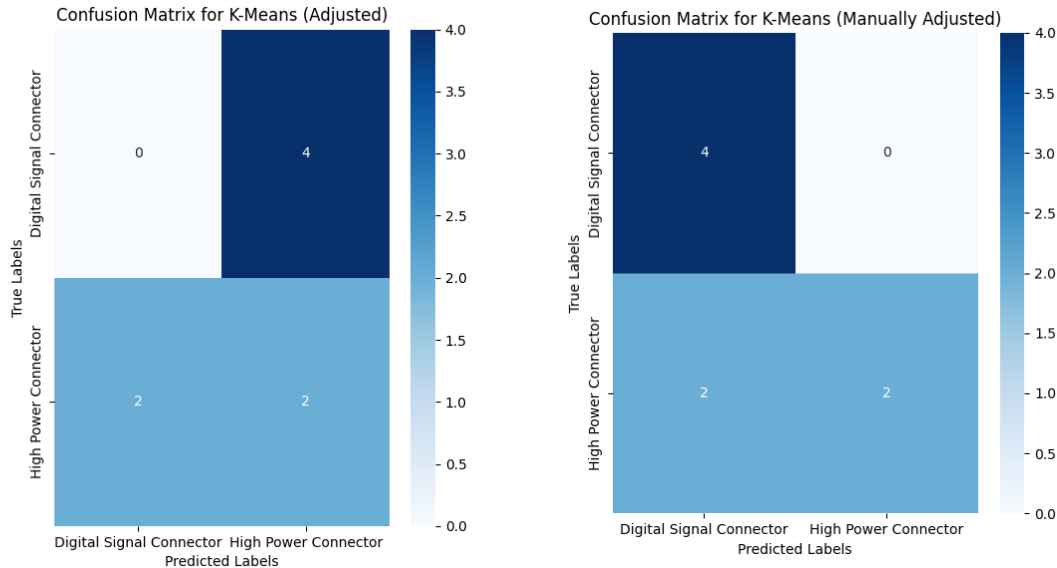


Figure 2.7: (a) confusion matrix before adjustment and (b) confusion matrix after adjustment

Figure 2.7 (a) illustrates that the labels in the clustering results are wrong and have to be switched. In other words, the two columns in this confusion matrix should be swapped to ensure the predicted labels align with the true labels. Swapping the labels manually for a small matrix like this is possible. The manually adjusted confusion matrix is shown in Figure 2.7 (b).

Still, for larger matrices, this will be too much manual work, and the Hungarian algorithm will optimally align the predicted labels with the true labels to maximize the correct prediction visualized on the confusion matrix's diagonal. This algorithm consists of four steps, where the first two steps are executed once, and steps three and four are performed iteratively until an optimal solution is found.

- The first step is to find the lowest value for each row and subtract the lowest value from each element in that row.
- Step two will subtract the lowest element per column and subtract that value from each element in that column.
- The third step covers all zeros in the matrix using a minimum of lines drawn through the rows and the columns. This step helps in identifying potential assignments.
- Step four includes creating additional zeros, which are unnecessary if the minimum number of covering lines equals the number of clusters because then an optimal assignment is found. If not, the smallest element not covered by a line in step three will be subtracted from all uncovered elements and added to all elements covered twice [42].

In the end, the optimal assignment is the same as the position of all the zeros in the adjusted matrix. This means that the position of each zero indicates that it is optimally paired with a predicted cluster to a true cluster [42].

To compare multiple large confusion matrices, a ratio derived from the matrices was used to summarize the performance of a clustering technique. The ratio is calculated as the percentage of the correct predictions relative to the incorrect predictions.

2.3.2 Testing

In the testing phase, new connectors were picked up and put in the dataset to test the performance of the clustering models on these connectors. The clustering models were evaluated on a new dataset to assess its generalizability. Also, the stability of the model was checked by running it multiple times to ensure the results were consistent.

Chapter 3

Results and discussion

3.1 Visualizing the data before using machine learning

Before using machine learning-based clustering to group the wire-harness connectors, various 2D and 3D plots were made to visualize some clusters potentially. This part of the thesis used a dataset of approximately 1,000 connectors without any true labels. For instance, total poles and the weight of connectors were used as the features for a 2D scatter plot. Normalization was performed to scale the data to a fixed range, between zero and one, using Min-Max scaling. In this case, the primary purpose is visualization; additional standardization is unnecessary. However, experimenting with standardization might be beneficial for clustering algorithms like K-means.

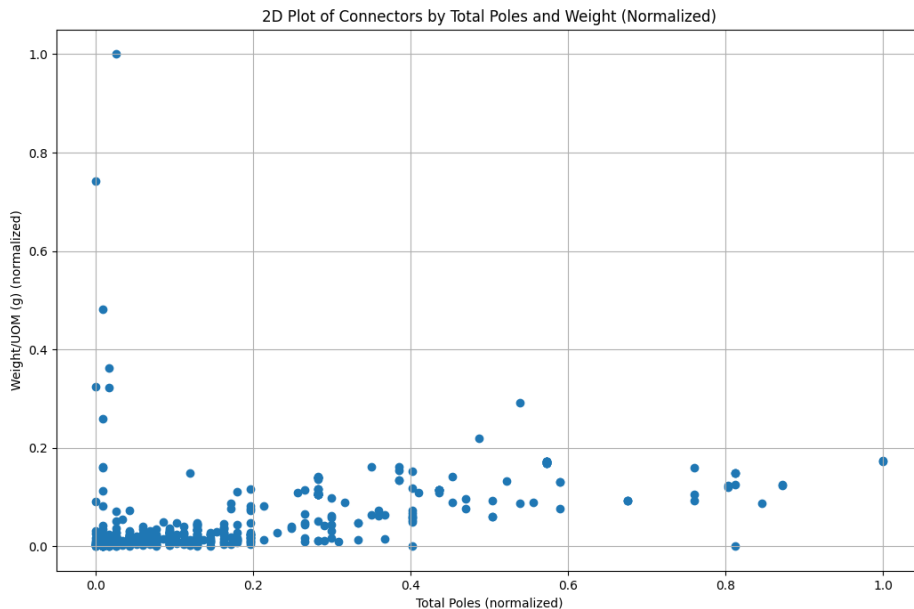


Figure 3.1: 2D plot of connectors based on total poles and weight

Figure 3.1 illustrates that there is a general trend indicating that if the number of total poles increases, the weight also tends to increase. Notably, there are outliers in the top left part of the

graph; these have a high weight but low total poles. These outliers correspond to high-voltage connectors, which need to handle high voltage safely due to their specialized build and materials.

To add another dimension to the analysis, the terminal size of the connectors was included. Similar to the 2D plot, normalization was performed to have the same scale for each feature.

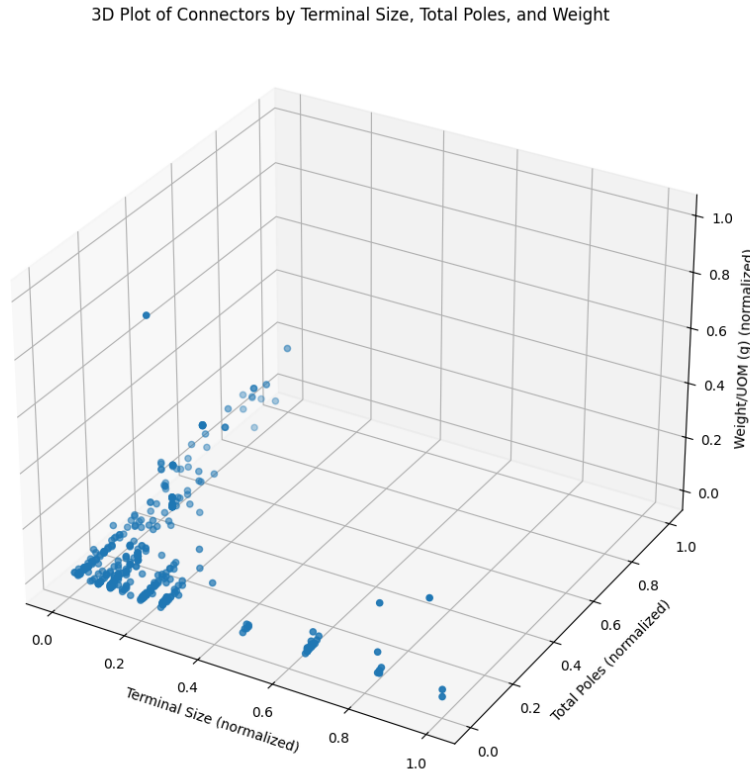


Figure 3.2: 3D plot of connectors based on terminal size, total poles, and weight

Figure 3.2 shows a noticeable trend where connectors with larger terminal sizes tend to have higher weights and more poles. Also, the plot shows a clear clustering of connectors with small terminal sizes, fewer poles, and lighter weight. Similar to the 2D plot, there are some outliers, especially connectors with larger terminal sizes but lower weights and total poles. They may represent specialized connectors that, despite their terminal size, need to be lightweight for specific applications.

3.2 Performing k-Means clustering on the 2D and 3D plot

The K-means algorithm was performed for the 2D and 3D plots shown earlier. The K-means algorithm requires a predefined number of clusters. The method used to find the optimal number of clusters is plotting the relationship between the number of clusters and the corresponding average silhouette scores [43].

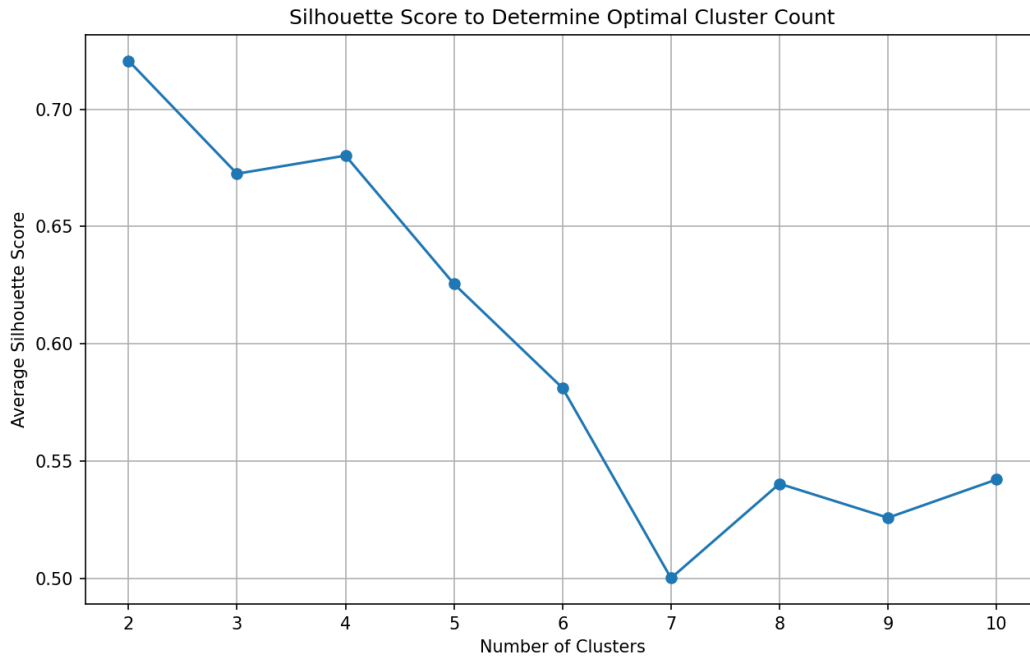


Figure 3.3: Silhouette score graph

Figure 3.3 demonstrates that the silhouette score is highest when the number of clusters is two, with a score of around 0.71. The graph indicates that the silhouette score significantly drops after four clusters. This suggests that two to four clusters are optimal for this dataset.

After knowing the optimal number of clusters, the K-means algorithm can be performed on the 2D plot. Each point in the graph represents a connector, and the color of the voxels indicates the cluster they belong to, determined by the K-means method.

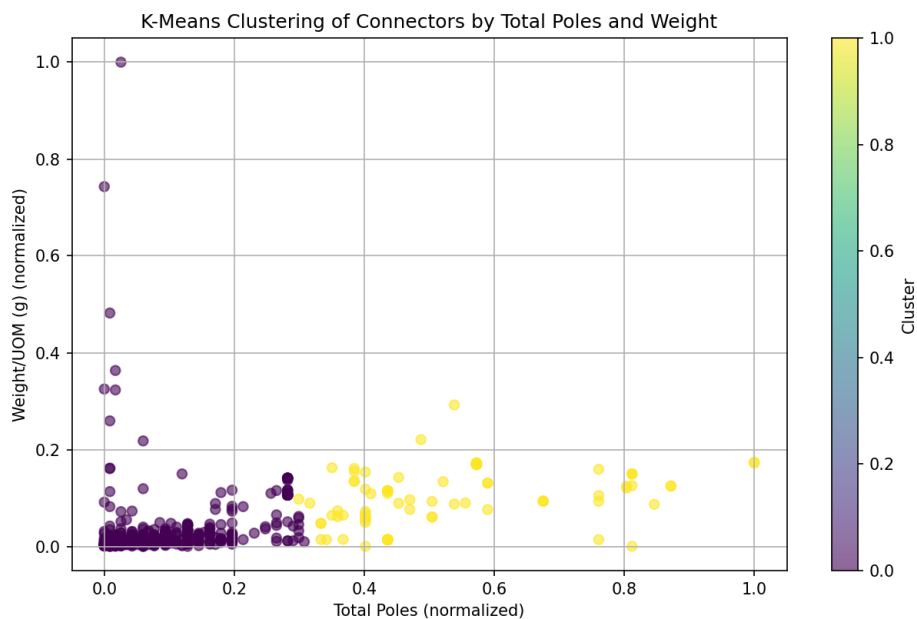


Figure 3.4: K-means clustering of connectors based on total poles and weight

Figure 3.4 shows two distinct clusters. The cluster represented in purple consists of connectors with a low amount of poles and weight, except for the outliers. In contrast, the yellow cluster includes connectors with a higher weight and amount of poles. These results show the importance of predicting the number of clusters and demonstrate the effectiveness of the K-means algorithm in identifying meaningful groups within the dataset.

A similar procedure is shown in Figure 3.5, where the average silhouette score is plotted against the number of clusters, this time for the 3D plot, including total poles, weight, and the terminal size of connectors.

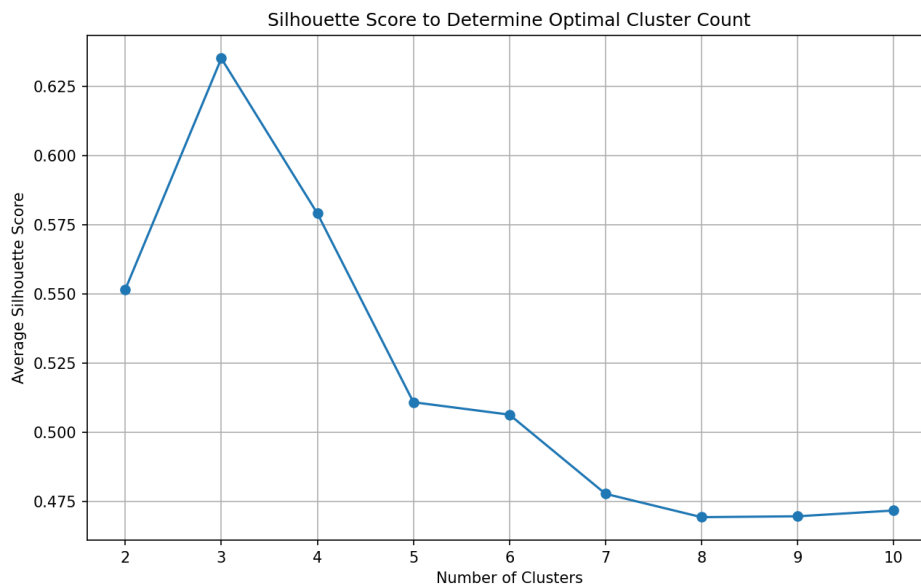


Figure 3.5: Silhouette score graph

Figure 3.5 illustrates that three clusters are the optimal number in this case, with a silhouette score of around 0.637.

The 3D plot shown in Figure 3.6 visualizes the connectors based on the three features, where the three colors, yellow, blue, and purple, represent the cluster to which each connector belongs, according to the K-Means clustering algorithm.

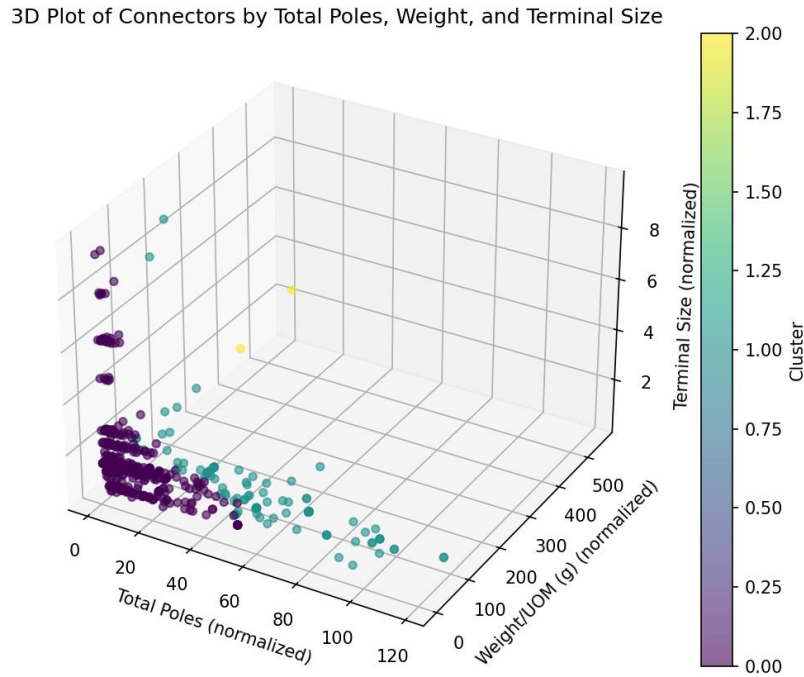


Figure 3.6: 3D plot of connectors after using K-means

From the plot, it is evident that most connectors are concentrated in the lower ranges of the axes. There are a few outliers with higher values, which are separated into different clusters. This distribution highlights the diversity within the dataset.

3.3 Unsupervised connector clustering

The results of different clustering methods performed on preprocessed data will be shown in this part of the results and discussion chapter. Before preprocessing, the dataset contained 17,776 connectors. However, after handling missing values, encoding categorical variables, feature scaling, handling outliers, dimensionality reduction, etc., the dataset decreased to approximately 6,500 connectors. This dataset did not contain any true labels, and the main objective of this part of the results is to explore the grouping of the connectors using different clustering methods, visualize the results, and find the differences between multiple clustering methods for this dataset.

3.3.1 K-means

For this dataset, before using the K-means and GMM methods, the first step is to do the silhouette analysis score again to check the best amount of predefined clusters. Figure 3.7 shows the graph, including the silhouette score per number of clusters.

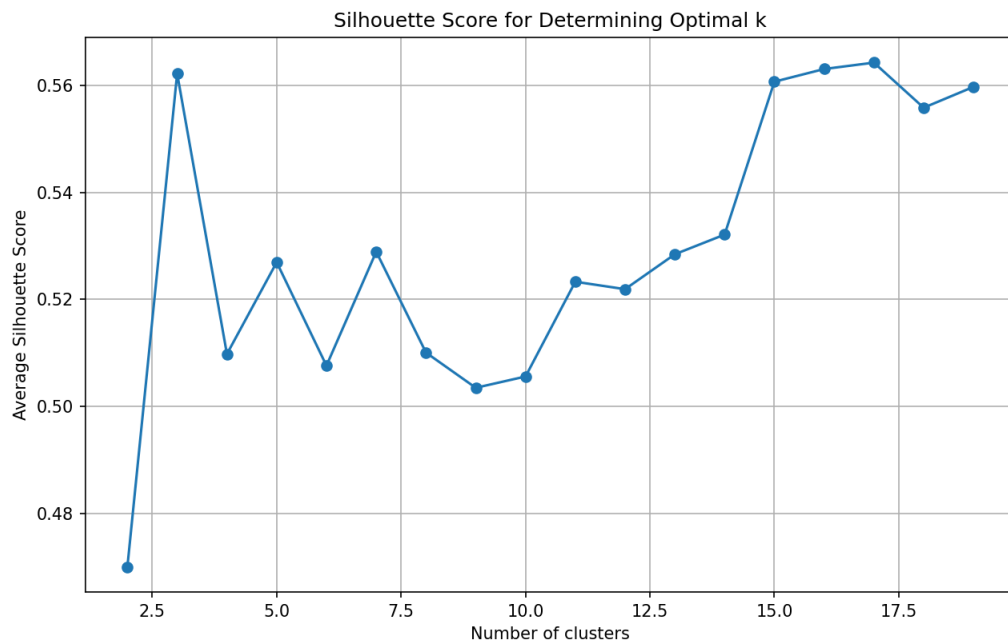


Figure 3.7: Silhouette score graph

The graph shows that the silhouette score increases to $k=3$, peaking at approximately 0.56. After three, the score decreases with minor fluctuations until $k=10$. After reaching the lowest point at $k=8$, the score increases again until a maximal score at $k=17$. This means this number of clusters will better represent homogeneous subgroups within the data. The number of 17 clusters is very interesting because later in the thesis, it will be discussed that this is indeed the correct amount of clusters, as with knowledge about the wire-harness connectors in Yazaki, it is known that there are 17 types of connectors.

Figure 3.8 depicts the result of applying the PCA method for visualizing the K-means clustering into 17 groups. The axes represent the first two principal components, which capture the largest variance in the data.

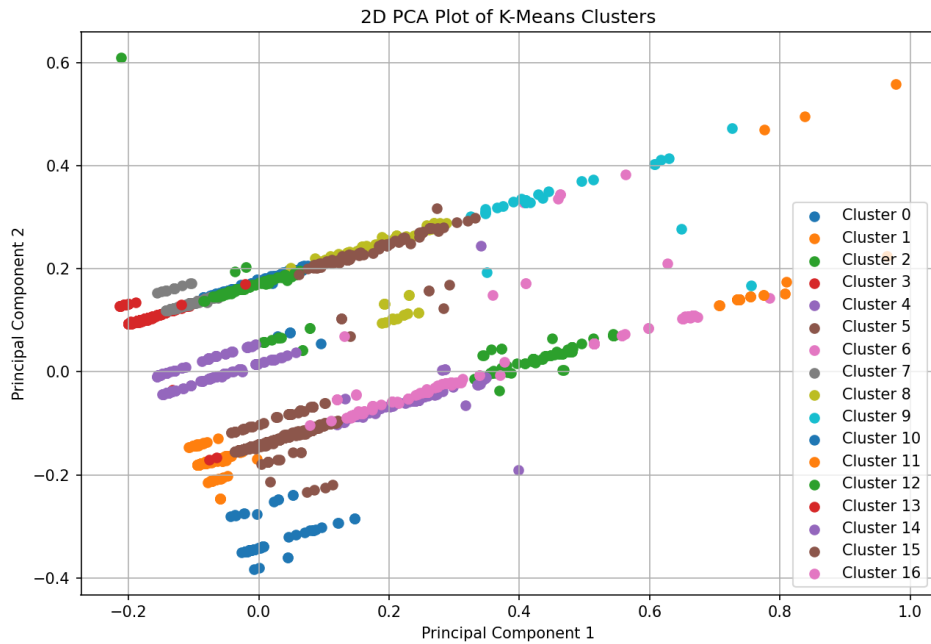


Figure 3.8: 2D PCA plot of K-means clustering

Figure 3.8 validates that the results of the K-means algorithm produce clusters of the same shape and size due to how the algorithm works. Because K-means doesn't handle noise by labeling it, it forces data points to clusters it may not belong to, which can misinterpret the results.

Figure 3.9 illustrates the clustering results using t-SNE, showing the 17 clusters, each with its unique color.

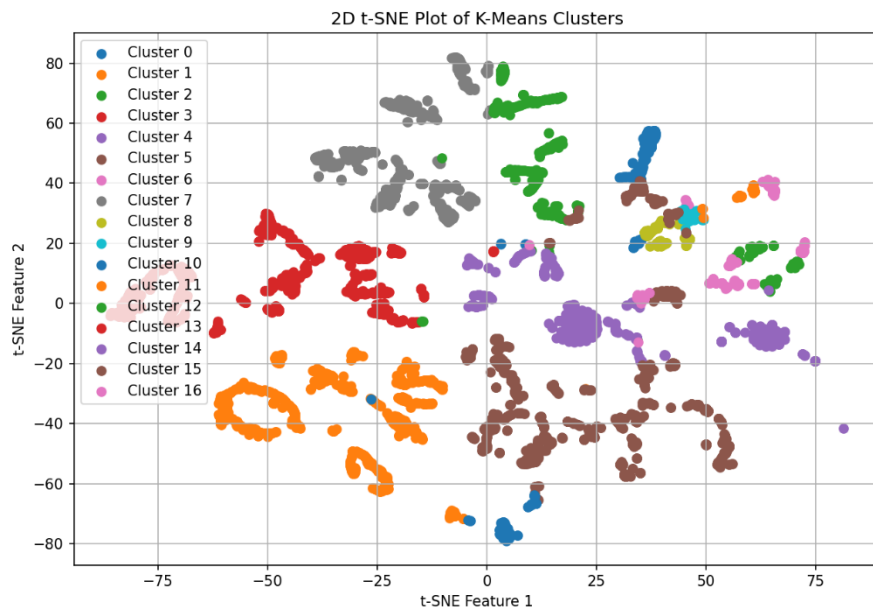


Figure 3.9: 2D t-SNE plot of K-means clustering

Using the PCA dimensionality reduction technique, the clusters are spread linearly along the principal components. Because this technique focuses on maximizing the amount of variance, the difference between the clusters is not as clear as with the t-SNE reduction technique. t-SNE is better suited for visualizing high-dimensional data by mapping similar points to nearby locations in 2D space. This visualizes groups that were not visible using the PCA technique.

Besides performing the K-means algorithm for 17 groups, it was also employed for three groups because of the high silhouette score of $k=3$ as well. Figure 3.10 shows the results for three groups using the t-SNE dimensionality reduction technique for visualization.

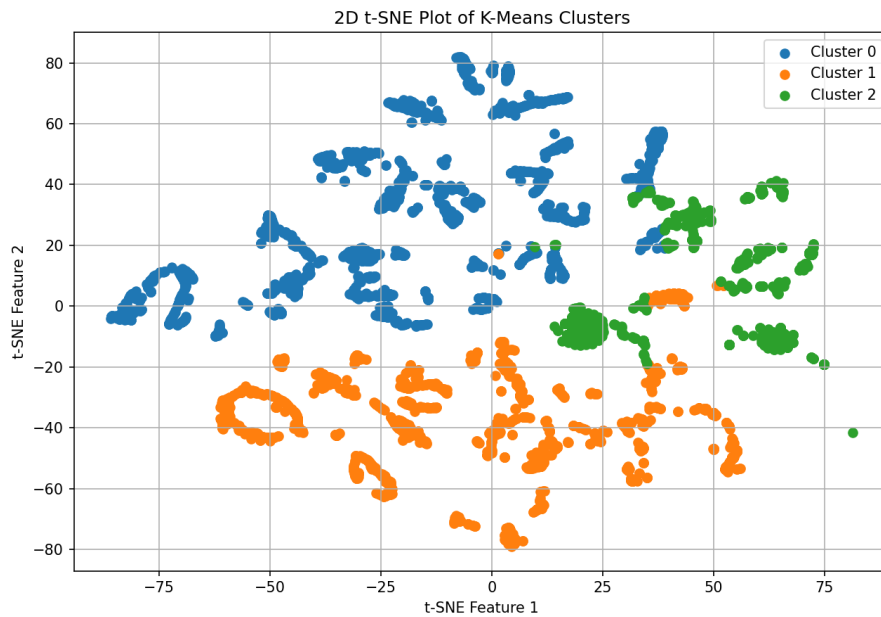


Figure 3.10: t-SNE plot for three groups

These clustering results were mapped back to the original feature space to draw a meaningful conclusion. To understand what each cluster represents, the average values of the original features were calculated. Using this procedure for the results of the K-means method when $k=3$, it can be generally stated that clusters zero and one focus more on simpler and standard connectors. In contrast, cluster two includes more complex connectors capable of handling more connections.

3.3.2 DBSCAN

Optimizing the parameters of the DBSCAN clustering method is a crucial step. These parameters determine the number of groups the dataset will be divided into. In this thesis, the epsilon and minimal sample features were manually adjusted. Another approach employed in this study involves finding optimal parameter values by calculating the average distances to the nearest points for the core point, plotting these distances, and identifying the point of maximum curvature. This method ensures the algorithm is finely tuned for the dataset.

When the epsilon is set at 0.95 and the minimal values are set at 5, the dataset is grouped into nine clusters using DBSCAN, as shown in Figure 3.11.

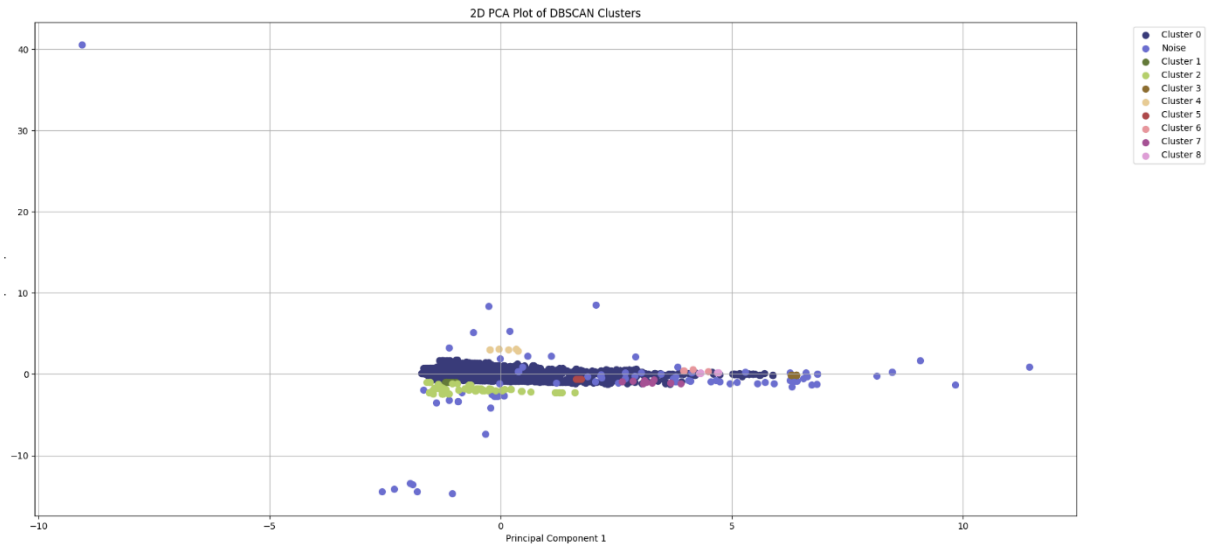


Figure 3.11: PCA plot of the DBSCAN results of 9 clusters

Figure 3.11 demonstrates that DBSCAN has identified nine clusters (cluster 0 to cluster 8) and a group called 'noise.' The noise points in light purple are spread across the plot, which means that they do not fit into the other clusters based on the parameters of this algorithm. Contrary to the K-means method, this does not force some data points to be grouped in a certain cluster. It is illustrated that most clusters appear elongated along the first Principal Component axis, which might indicate that the clusters would be more complex in the original feature space. The difference between the plots of the K-means and DBSCAN results is the clusters' shape and size. Contrary to the similar size and shape in the results of K-means, it identifies clusters of different shapes and densities, which can be seen in how the data points are spread in Figure 3.11. This gives more insights into datasets with more complex structures and high dimensions.

After tuning the parameters and setting epsilon at 0.65 and the minimal samples parameter at 17, the dataset of around 6,500 connectors was grouped into 17 clusters using DBSCAN.

3.3.3 GMM

The third method for grouping the connector dataset is the GMM clustering method. Figure 3.12 demonstrates the results of that method, visualized by a 2D PCA plot.

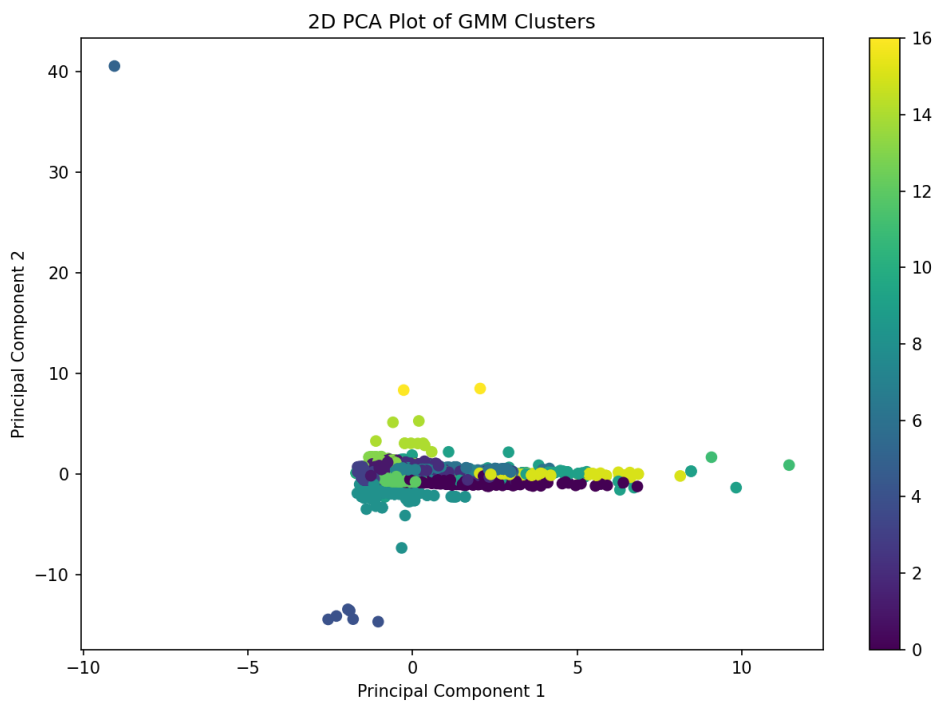


Figure 3.12: PCA plot of GMM results of 17 clusters

It can be seen that most of the data points lie in the center of the plot. The clusters are more compact compared to the results of K-means and DBSCAN.

The same clustering method was also performed to group the dataset into three groups, as shown in Figure 3.13.

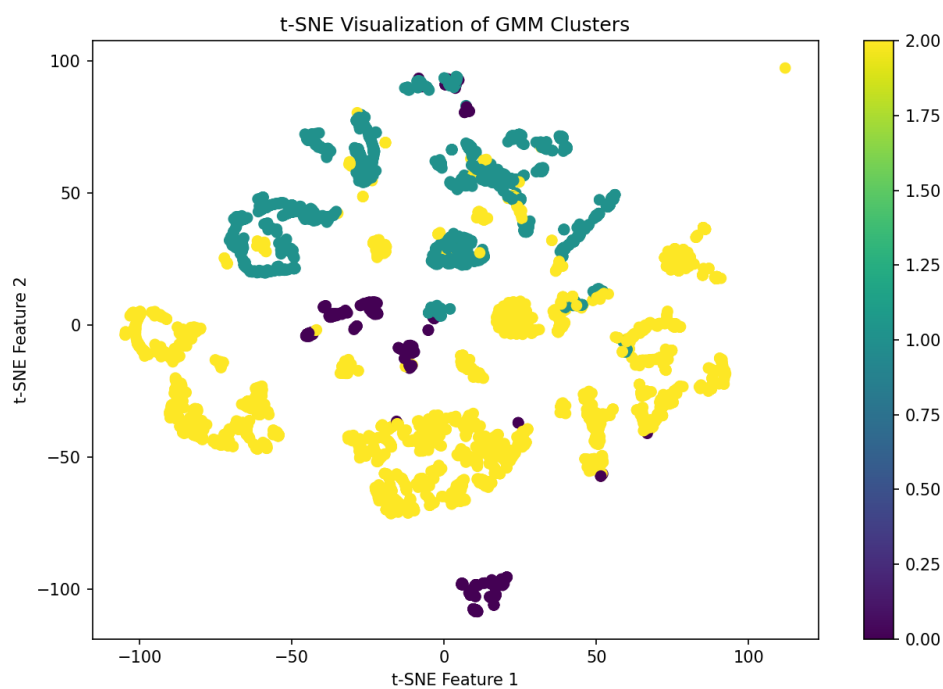


Figure 3.13: t-SNE plot of GMM results of three clusters

The t-SNE plot gives more visual clarity for the results than the PCA plot. Once more, Figure 3.13 illustrates that GMM groups the data more compactly. This method does not handle the outliers.

The GMM method works by assigning a probability to a data point belonging to a cluster; this offers a more nuanced understanding of the grouping than K-means and DBSCAN. The disadvantage of assigning a probability to each data point is that the computational requirements are more intense than the K-means method.

3.3.4 Decision tree

The K-means method does not provide additional insights into why certain data points are grouped together. Decision trees can be used to make it explainable by identifying the key features that differentiate clusters. It tries to capture the logic, and by visualizing the tree, it can be seen. To keep the decision tree clear, the K-means algorithm was chosen to be performed on a 3D plot containing the minimum and maximum temperature and the total poles, which is shown in Figure 3.14.

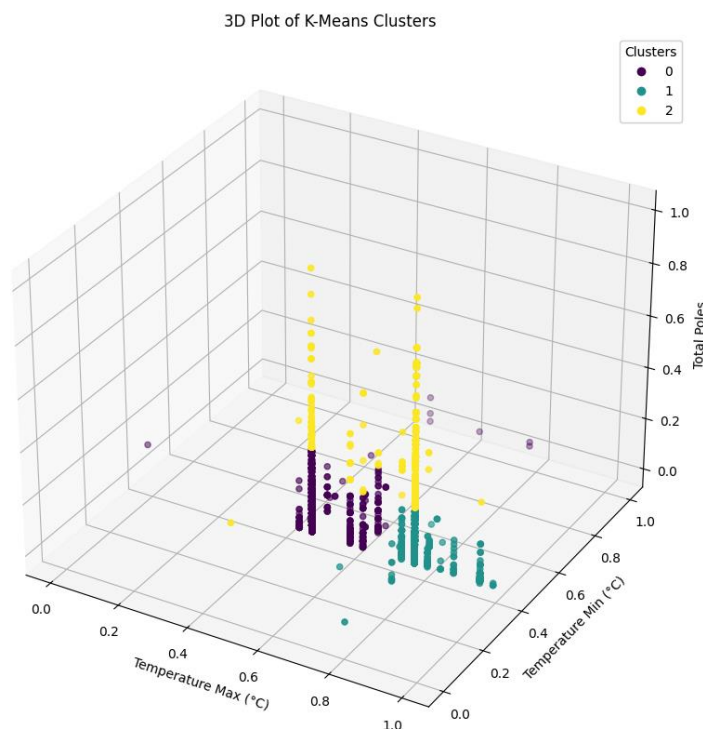


Figure 3.14: 3D plot of K-means clusters

It can be determined that the dataset is grouped into three clusters, represented by different colors: cluster zero in purple, cluster one in blue, and cluster two in yellow. It can be stated that

cluster one has a high density of points for higher values of the maximal temperature, indicating that these connectors are probably made for high-temperature applications. On the other hand, cluster two is more spread in all three dimensions and could represent connectors used for diverse applications. Finally, cluster zero has a high density of points for lower values temperature values.

Based on these results, a decision tree is made for explainability, as shown in Figure 3.15.

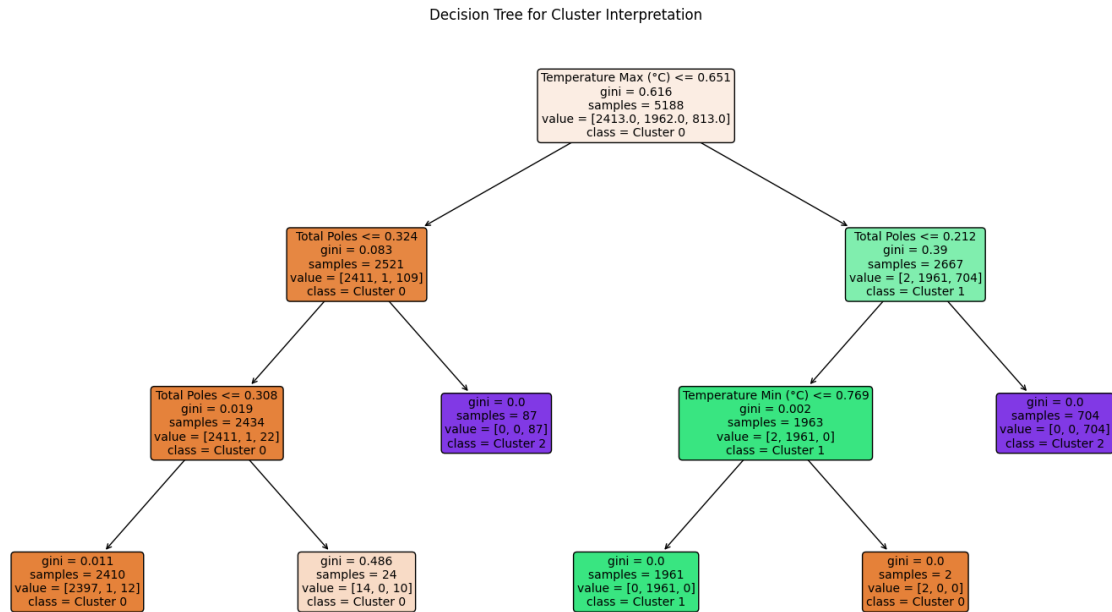


Figure 3.15: Decision tree based on the K-means results

Each node or rectangle in this decision tree represents a decision and includes one of the three features. The leaf nodes, the rectangles at the bottom of the three, and no more arrows from depart are the eventual clusters in which the connector will be grouped.

An example interpretation of the decision tree will be made to explain the working. For example, looking at the first node or root node, the data points will be moved to the left subtree if the maximal temperature is less or equal to the normalized value of 0.651. If the connector has an amount of normalized total poles lower or equal to 0.324, it will be checked if the normalized total poles are lower or equal to 0.308, and if that is true, the data points belong to cluster zero.

There are also other values in the nodes that give extra information about decision-making. Gini is the value that represents the impurity of the node; if this value is zero, then that means that this node is homogeneous. The number of samples or data points inside the nodes is also given, and the value represents the distribution of the data points over the different clusters. Finally, the class indicates the cluster to which most of the data points belong.

It can be concluded that the decision tree offers a clear visual interpretation of the decisions made in the clustering process. Next, compared to the K-means method, the decision tree is less computationally hard because of its simplicity.

3.4 Clustering with true labels and comparative analysis

This section of the thesis represents using different machine-learning clustering methods to group the wire-harness connectors applied to a dataset containing 1,020 connectors enriched with true labels. There are 17 connector types in total, and the dataset contains 60 connectors per type. Thus, besides the characteristics per connector, the connector type is also included in the dataset. Because clustering is normally done to find groups in a dataset without true labels, this new dataset provides a unique opportunity to validate and refine the clustering techniques and suggest an approach specifically for grouping Yazaki's wire-harness connectors. The performance of a specific method was measured using confusion matrices and silhouette scores. This thorough analysis makes the approach robust and validated for the grouping of connectors.

Five comparisons were made to find the best possible approach for the connectors: with or without standardization, with or without adding weights to the characteristics based on their importance, different initialization techniques, balanced K-means compared to normal K-means, and with or without handling outliers. Doing these five comparisons, it was chosen to work with the K-means method. For all of the comparisons, the same preprocessed dataset is used.

3.4.1 Standardization

This part of the thesis compares using K-means on the preprocessed dataset of 1,020 connectors with true labels and with or without standardization.

With standardization

First, using K-means clustering to group the connectors, the numerical data was standardized. Figure 3.16 illustrates the PCA plot with the results using standardization.

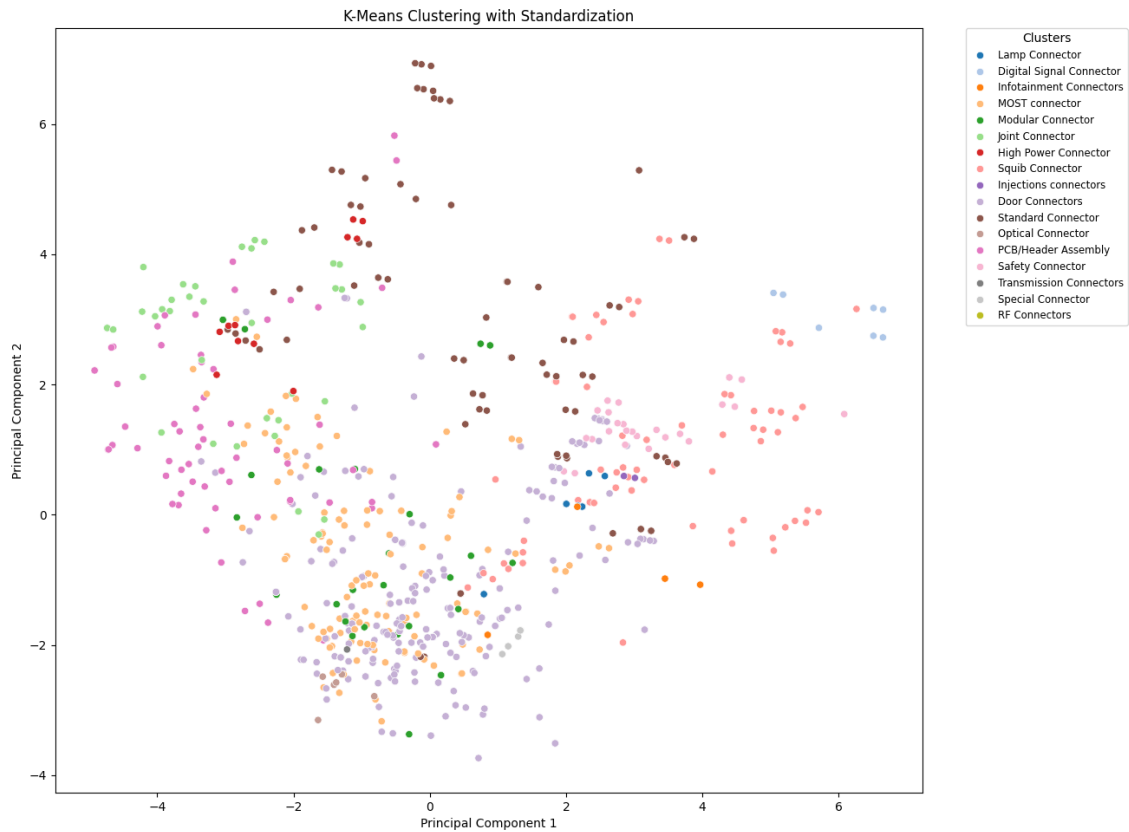


Figure 3.16: K-means clustering with standardization

Figure 3.17 demonstrates the confusion matrix of these results to validate the K-means and compare this method to others.

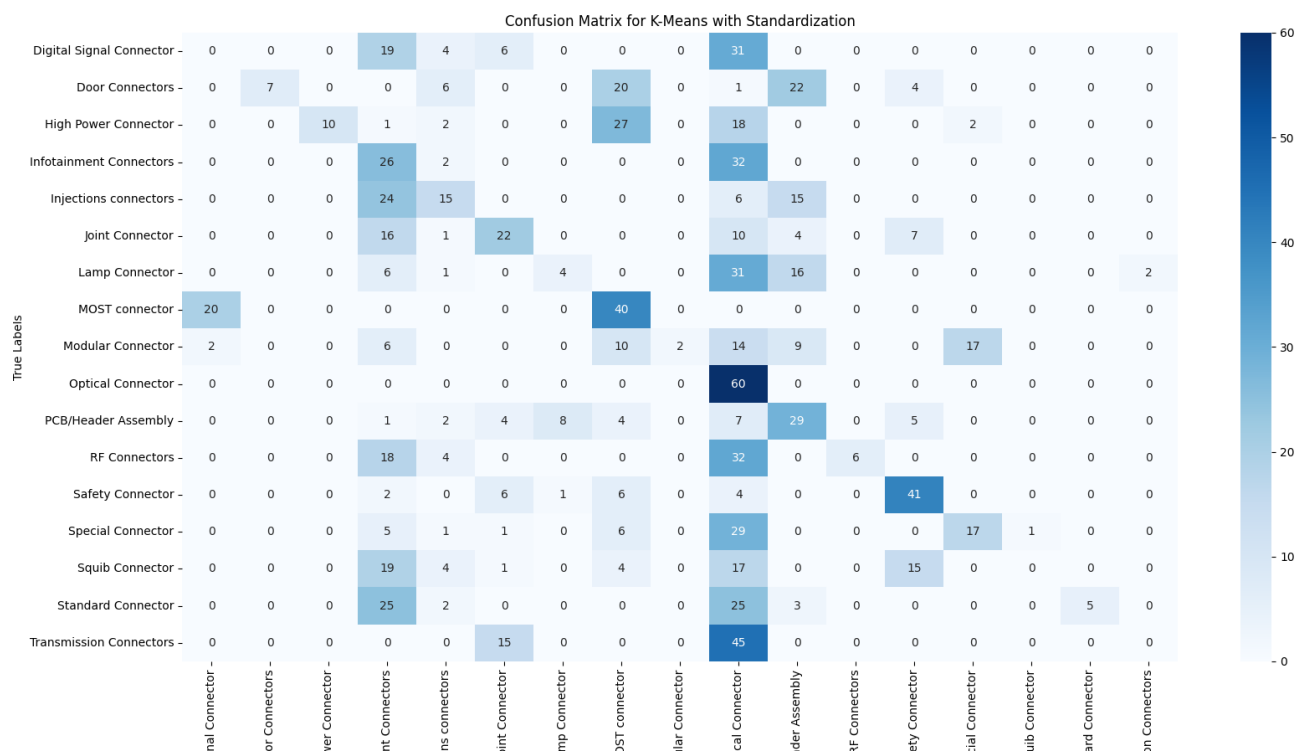


Figure 3.17: Confusion matrix with standardization

The original dataset of 1,020 connectors consisted of 60 connectors per type; after clustering, the grouped connectors are shown. The matrix's rows are the connectors' true labels, and the columns represent the predicted labels assigned using the K-means algorithm. The diagonal values in the matrix represent the correctly grouped connectors, and the values off-diagonal represent the wrongly classified connectors. According to Figure 3.17, the optimal connectors are grouped perfectly, and another type that was grouped well is the MOST connectors, for example, with 40 correctly clustered instances.

The ratio of the correct predictions relative to the incorrect predictions in percentage here is 38.59%. Also, the silhouette score for K-means with standardization is 0.067.

Without standardization

The K-means algorithm was performed on the dataset of 1,020 connectors with true labels without standardizing the numerical values. The results are illustrated in a 2D PCA plot, as shown in Figure 3.18.

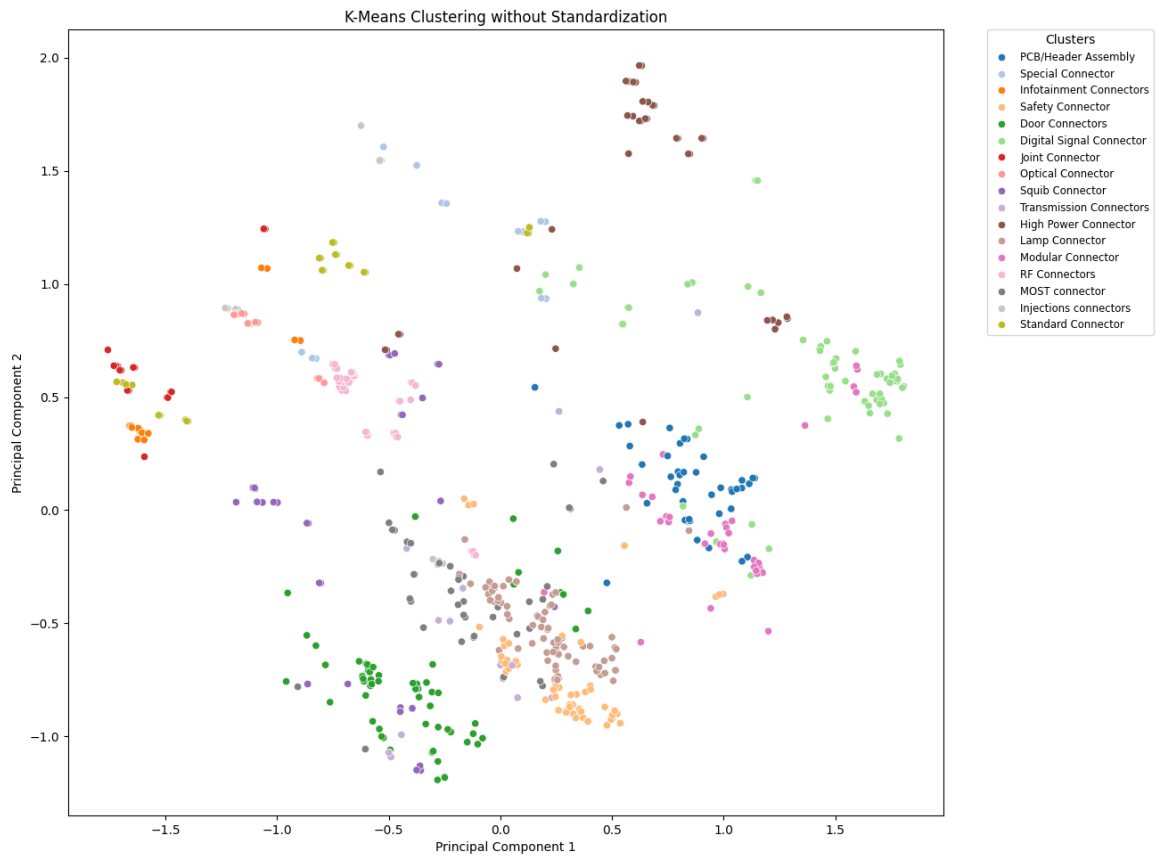


Figure 3.18: K-means clustering without standardization

Figure 3.19 shows the confusion matrix based on the clustering results using K-means without standardization.

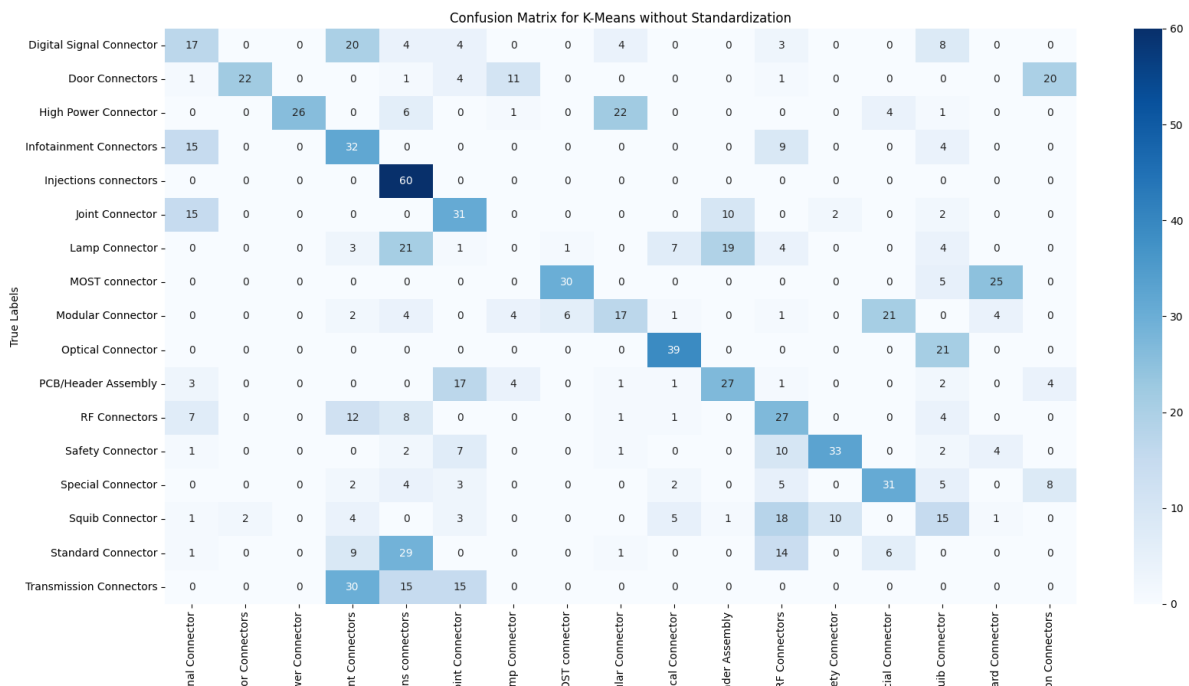


Figure 3.19: Confusion matrix without standardization

The ratio of the correct predictions relative to the incorrect predictions in percentage here is 66.39%. Also, the silhouette score for K-means with standardization is 0.180.

Comparison and discussion

The clustering results with or without standardization indicate that the K-means algorithm performed better without standardization, with a difference of 27.8% in terms of the ratio in percentage and a difference of 0.113 in terms of the silhouette score. A higher silhouette score means better-separated clusters.

Thus, it can be concluded that it is not necessary to standardize the numerical values because it decreases the performance of the clustering method. The numerical values were already normalized and, because of this, already have the same scale. Data standardization before clustering is commonly done because clustering models are distance-based models and standardization ensures that the high ranges will not have a bigger influence on clustering [44]. However, if the data is standardized, a big assumption is made about how it is distributed, as it is best for normal distributions, and it is not sure if the numerical values are distributed like that [44].

3.4.2 Adding weights to the characteristics

After knowing that performing the K-means algorithm to cluster connectors yields better results when not standardizing the numerical values, from now on, standardization will not be used anymore for the rest of the comparisons in this part of the thesis.

After normalizing the numerical values and one-hot encoding the categorical values of the dataset, it is possible to add weights to those values based on their importance. Research was done at Yazaki to understand which connector characteristics are most important in terms of the interchangeability of the connectors. The characteristics of gender, hybrid poles, pin rows, and terminal size were most important. Thus, their values were multiplied by five before clustering. The characteristics of the cavity family and maximal temperature were also more important than the not-mentioned characteristics but not as important as the ones multiplied by five, so they were multiplied by three.

Adding weights to the most important characteristics

First, using K-means clustering to group the connectors, a weight was added to the most important connector characteristics. Figure 3.20 illustrates the PCA plot with the results using the weights method.

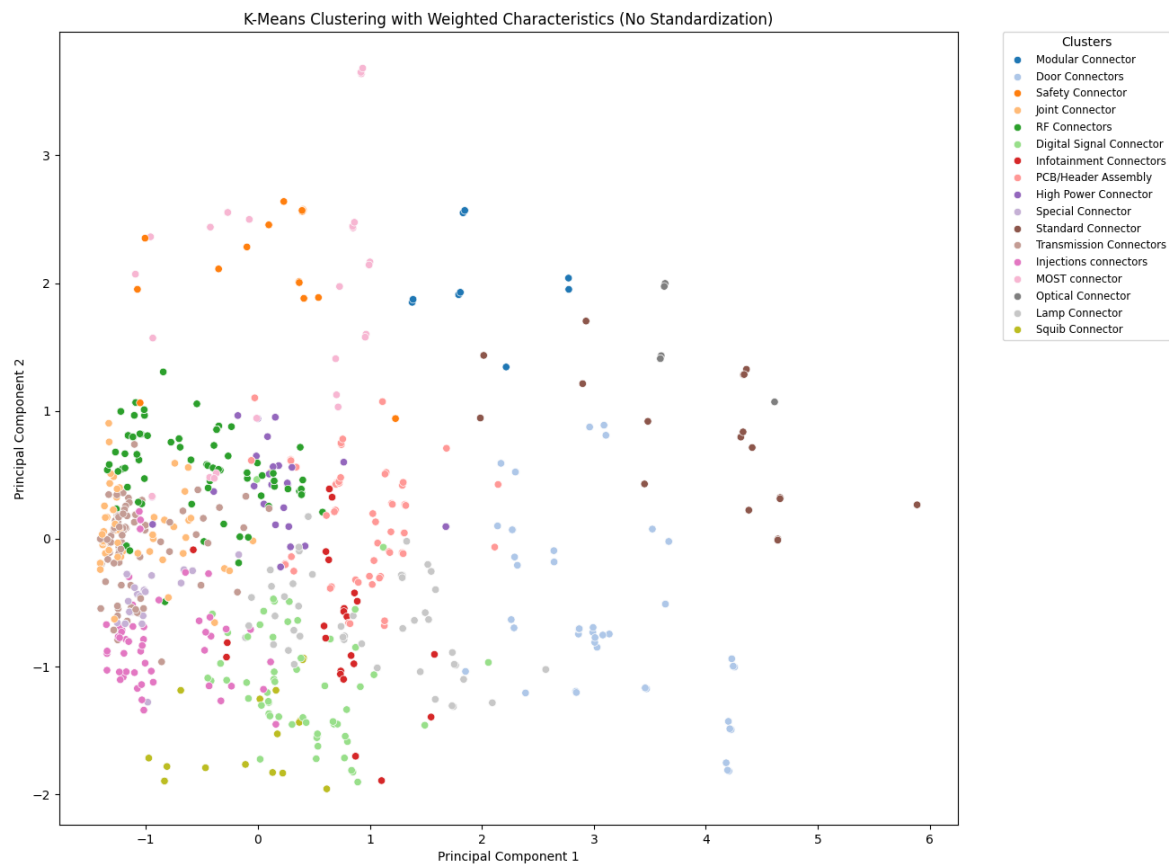


Figure 3.20: K-means clustering with weighted characteristics

Figure 3.21 shows the confusion matrix based on the clustering results using K-means with weights added to the most important characteristics.

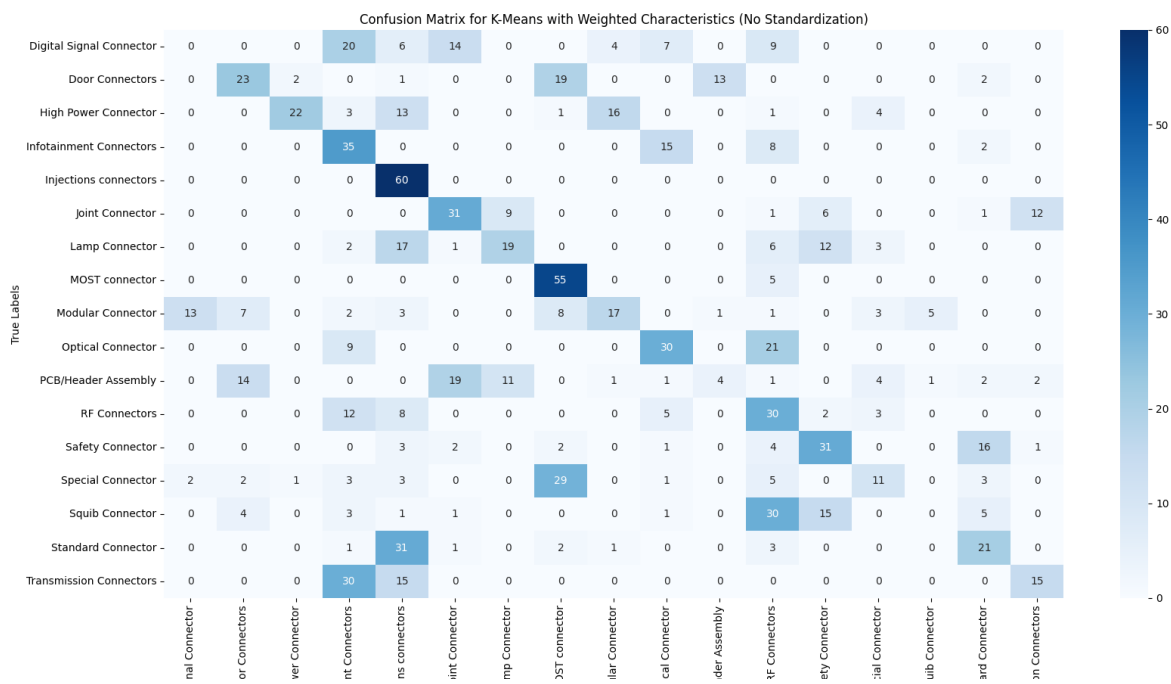


Figure 3.21: Confusion matrix with weighted characteristics

The ratio of the correct predictions relative to the incorrect predictions in percentage here is 65.58%. Also, the silhouette score is 0.153.

Without adding weights to the most important characteristics

The results of performing the K-means method without adding the weights are the same as those of the K-means without standardization.

The percentage ratio of the correct predictions relative to the incorrect predictions is 66.39%. Also, the silhouette score is 0.180.

Comparison and discussion

It can be stated that the method without adding weights to the most important connector characteristics yields better results. It has a higher ratio of 0.81 percent and an increase of the silhouette score of 0.027. Although the most important features had a higher impact on the clustering, it didn't yield better results.

3.4.3 Different initialization techniques

After knowing that performing the K-means algorithm to cluster connectors yields better results when not standardizing the numerical values and not adding weights to the most important features, from now on, standardization and adding weights will not be used anymore for the rest of the comparisons in this part of the thesis.

The K-means algorithm is sensitive to the initialization of the first centroids. The clustering results will be suboptimal if the first centroids are placed poorly. This is why two different initialization techniques were compared: picking up one connector per group for the first centroid and performing multiple random initializations while selecting the solution with the best results.

Using one connector per group for the first centroid

Because the dataset used in this part of the thesis includes true labels, it is possible to adapt the initialization of the K-means algorithm specifically for wire-harness clusters by selecting one connector per group as the first centroid. This way, there is certainty that it is not random and is selected well, based on the information available in the dataset. Figure 3.22 illustrates the PCA plot with the results using one connector per group as the first centroid. The connector picked per group is a connector with the average of every feature, so it is ensured that the picked connector represents that group best.

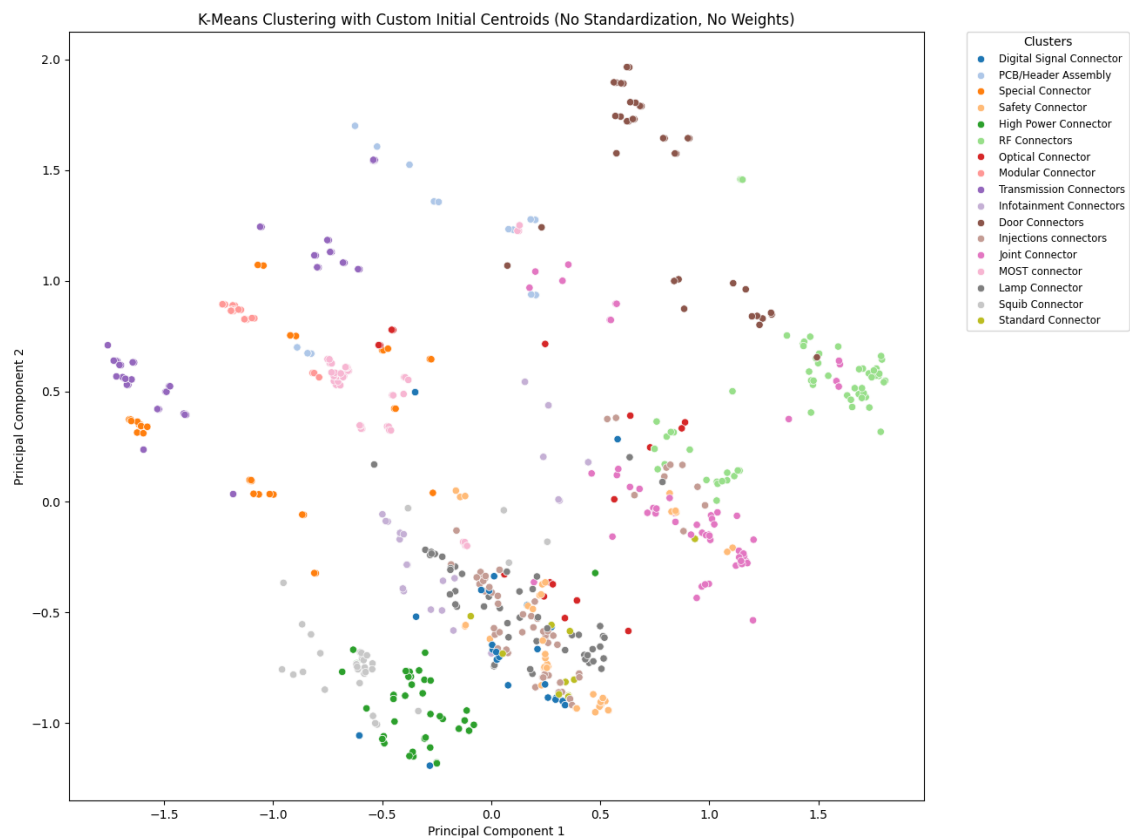


Figure 3.22: K-means with custom initial centroids

Figure 3.23 shows the confusion matrix based on the K-means clustering results while using one connector per group as the first centroid.

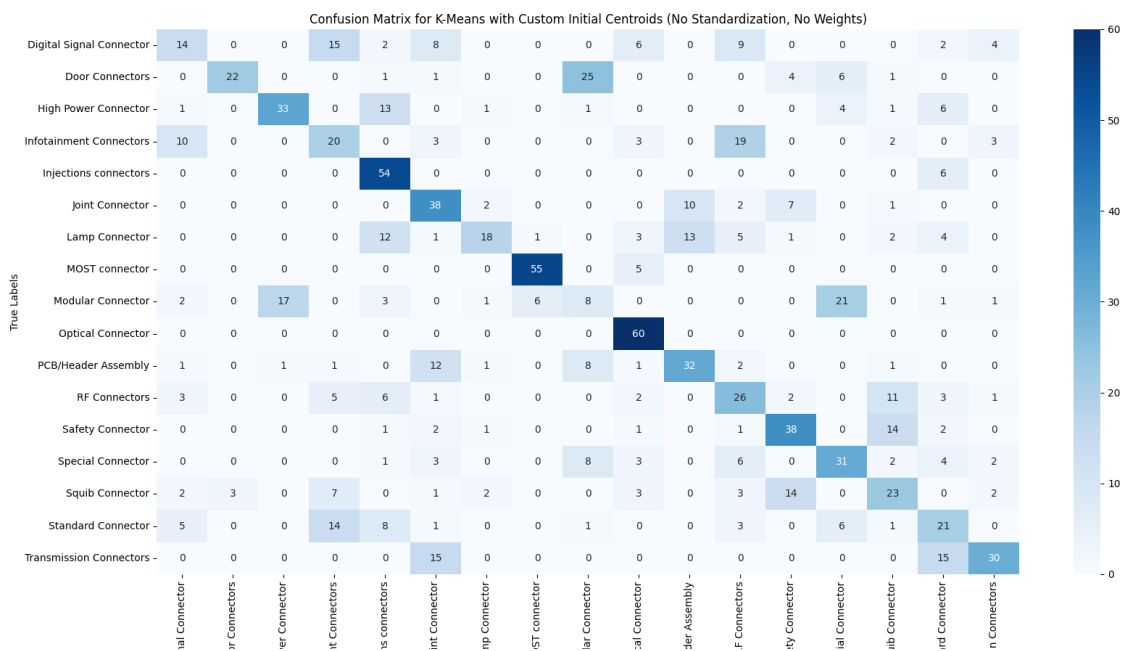


Figure 3.23: Confusion matrix with custom initial centroids

The ratio of the correct predictions relative to the incorrect predictions in percentage here is 105.23%. Also, the silhouette is 0.163.

Multiple runs of random initialization

The results of performing the K-means method using multiple runs of random initialization are the same as the clustering results without adding additional weights because, here, the standard initialization method is used in this thesis. The K-means algorithm uses random initialization and runs it ten times, and the one with the lowest variance inside the clusters is picked.

The percentage ratio of the correct predictions relative to the incorrect predictions is 66.39%. Also, the silhouette score is 0.180.

Comparison and discussion

It can be stated that the method of using the average values of the connectors as the first centroid yields better results in terms of ratio; namely, it increased the ratio by 38,84%. On the other hand, this initialization method decreased the silhouette score, namely 0.017, compared to the initialization method performing multiple runs with random initialization.

The significant increase in the ratio indicates that using a connector as the first centroid improves the correctness of the predictions because this method uses domain knowledge, which ensures that each first centroid is a valid representation of a connector type. However, this method had a slight decrease in terms of the silhouette score, which indicates that the clusters were less separated. A potential reason could be that the connectors as the initial centroids may not be optimal in terms of distance from each other. Next, running the initialization multiple times makes the method more computationally intensive.

3.4.4 Balanced K-means

After knowing that performing the K-means algorithm to cluster connectors yields better results when not standardizing the numerical values, not adding weights to the most important features, and using the mean values of a group of connectors as the initial centroid, from now on, these methods will be used for the rest of the comparisons in this part of the thesis.

In this study, it was checked if performing balanced K-means would improve the clustering performance. The balanced K-means clustering method requires all clusters to be the same size.

Performing balanced K-means

Figure 3.24 demonstrates the PCA plot with the results after using the balanced K-means method.

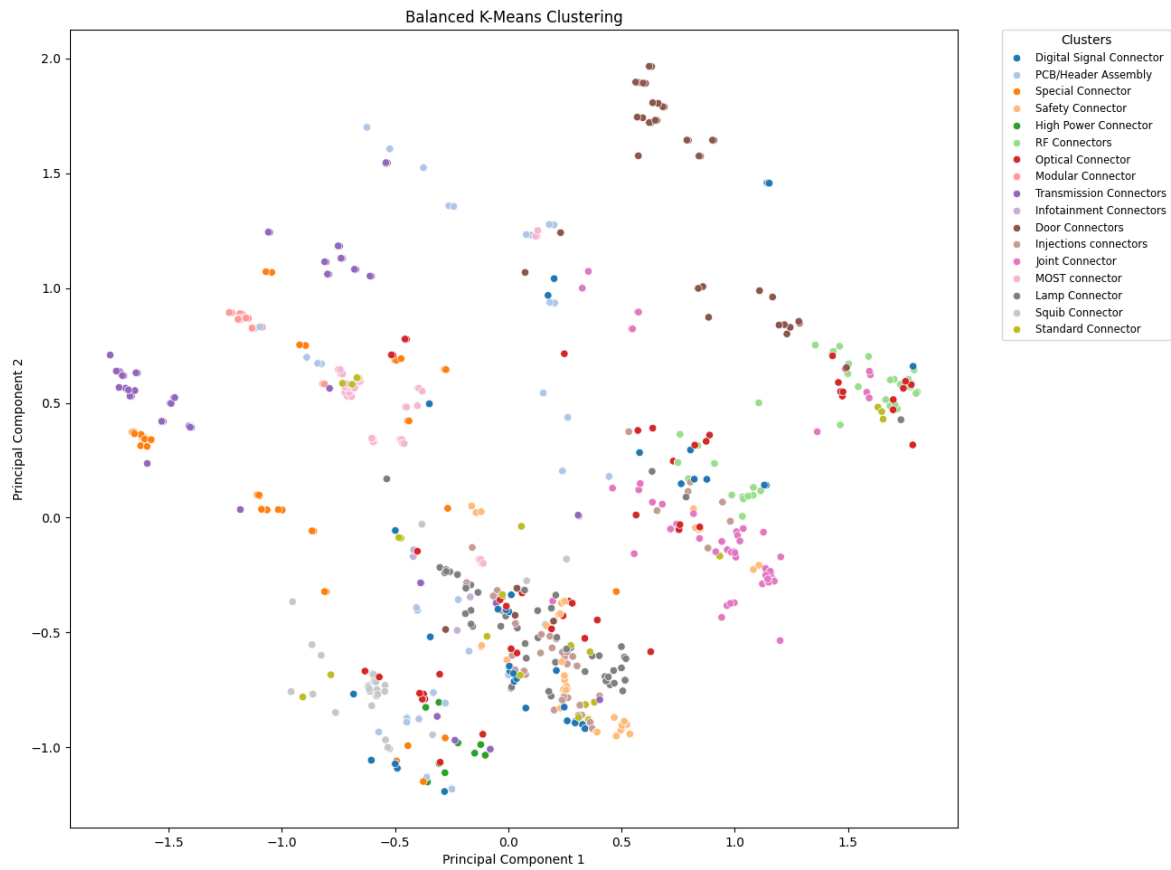


Figure 3.24: Balanced K-means

Figure 3.25 shows the confusion matrix based on the balanced K-means clustering results.

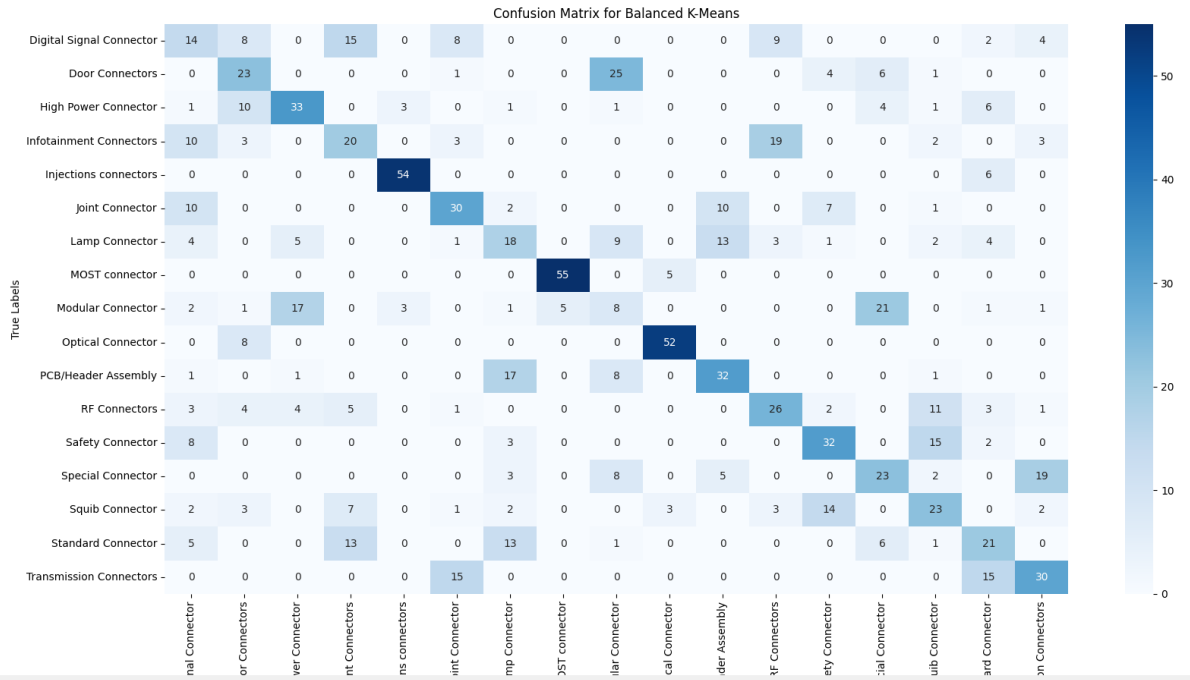


Figure 3.25: Confusion matrix of balanced K-means

The ratio of the correct predictions relative to the incorrect predictions in percentage here is 93.92%. Also, the silhouette score for K-means with standardization is 0.108.

Performing normal K-means

The results of the K-means method are the same as those in the results of using one connector per group as the initial centroid.

The ratio of the correct predictions relative to the incorrect predictions in percentage here is 105.23%. Also, the silhouette score is 0.163.

Comparison and discussion

Compared to the normal K-means method, the balanced K-means performance decreased. The ratio of the correct predictions relative to the incorrect predictions lowered by 11.31%, and the silhouette score lowered by 0.055.

The restriction of the balanced K-means method is that it forces every cluster to have the same amount of data points. Because of this, connectors from the correct cluster might be put in another cluster to preserve equilibrium, which may result in abnormal cluster borders.

3.4.5 Handling outliers

The final part of comparing different ways of performing K-means and improving the clustering technique specifically for wire-harness connectors includes handling outliers. It was chosen to remove 20% of the rows with the highest number of average outliers. 80% of the dataset was kept; thus, approximately 800 connectors with true labels were kept.

Figure 3.26 demonstrates the PCA plot with the results after handling the outliers of the dataset.

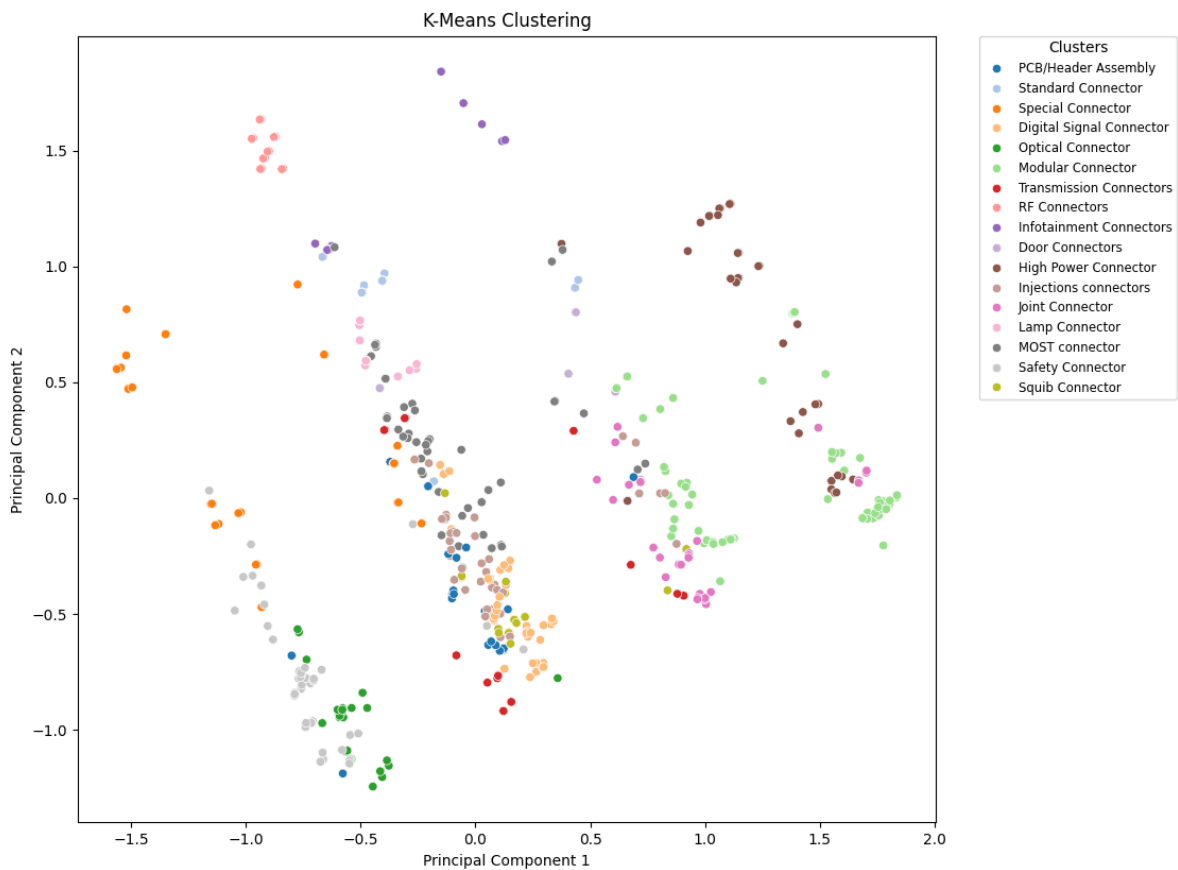


Figure 3.26: K-means clustering after handling outliers

Figure 3.27 shows the confusion matrix after handling the outliers.

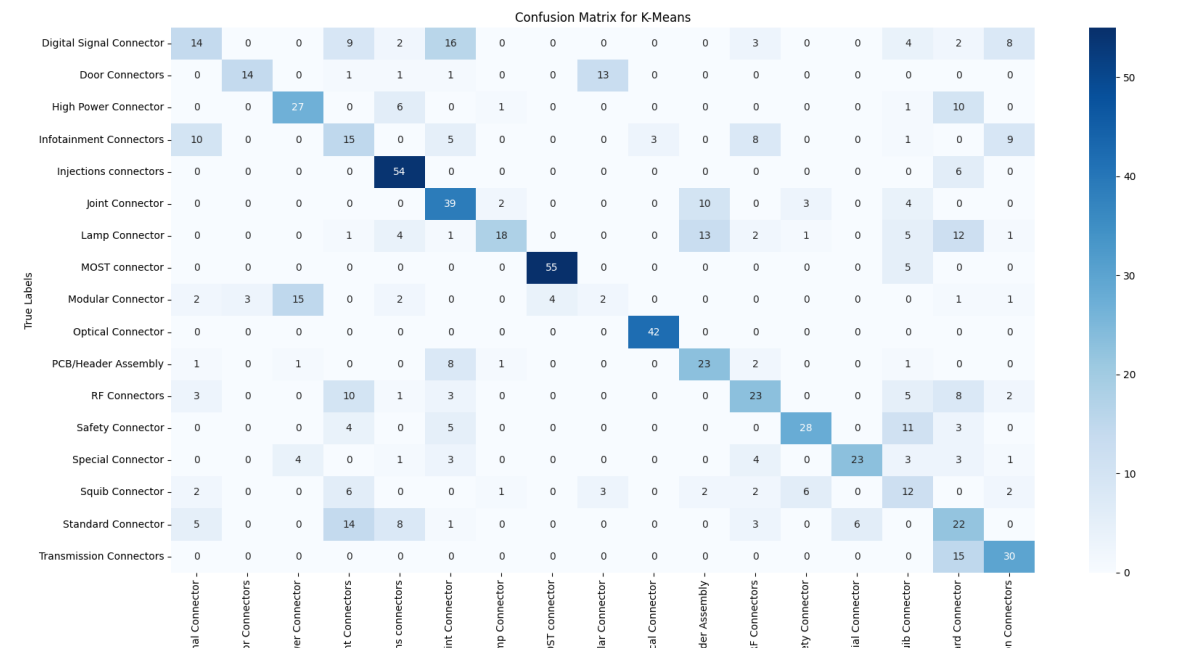


Figure 3.27: Confusion matrix after handling outliers

Figure 3.27 demonstrates that the MOST connector shows the highest level of correctly grouped connectors, namely 55. Injection connectors have the second highest number of correctly grouped connectors, namely 54. However, injection connectors have 14 more misclassified connectors into other types compared to MOST connectors. Digital, door, and high-power connectors have the lowest amount of correct predictions.

The ratio of the correct predictions relative to the incorrect predictions in percentage here is 117.29%. Also, the silhouette score is 0.203. The accuracy of this method is 62%.

Comparison and discussion

After handling the outliers, the ratio of the correct predictions relative to the incorrect predictions increased by 12.06%, and the silhouette score increased by 40 compared to using the K-means method without handling outliers.

This means that using the K-means method, without standardizing the data, without adding weights, using the average values of the connectors per type as the initial center, and handling the outliers gives the best performance of K-means in terms of the confusion matrices and the silhouette scores specifically for grouping the wire-harness connectors.

3.5 Most similar wire-harness connector

To find interchangeable connectors, in this thesis, it was chosen to pick up only the MOST connector group after the clustering result because of its highest number of correctly grouped connectors and group it again using the DBSCAN algorithm.

Because the DBSCAN algorithm requires two parameters, a k-distance graph was made to show the sorted distances of the k-th nearest neighbor for each point in the dataset. Figure 3.28 illustrates the k-distance graph for the MOST connector group.

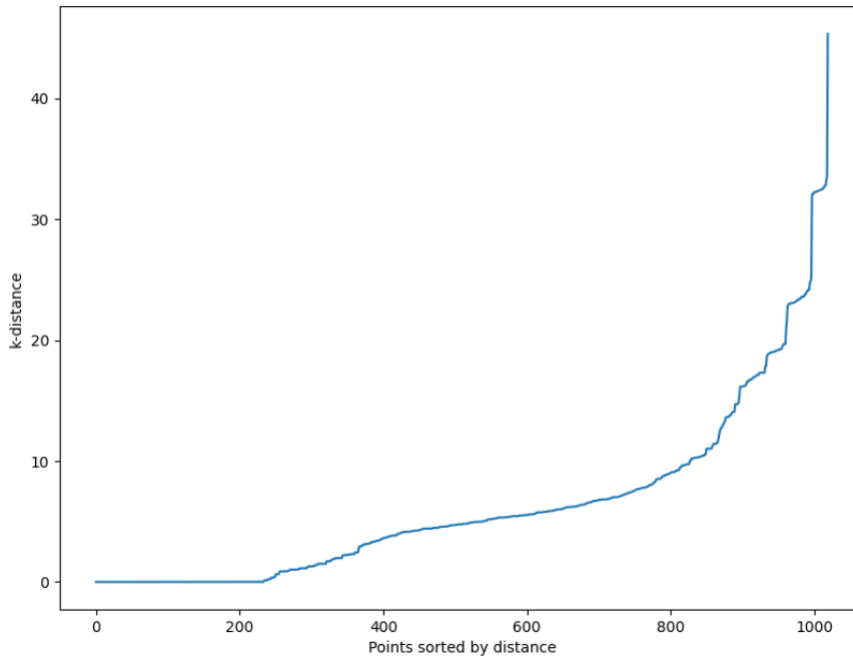


Figure 3.28: K-distance graph

The rule for the k-distance graph is to look for the point where there is a significant change in the slope of the curve. This point is called the elbow. The epsilon parameter should be chosen at the point where the curve transitions from a gentle slope to a steep slope. In this case, an epsilon around ten would be the best option [45].

Using ten as the value for epsilon, the DBSCAN clustering method grouped the MOST connector group into ten clusters (cluster 0 to cluster 9), as shown in Figure 3.28.

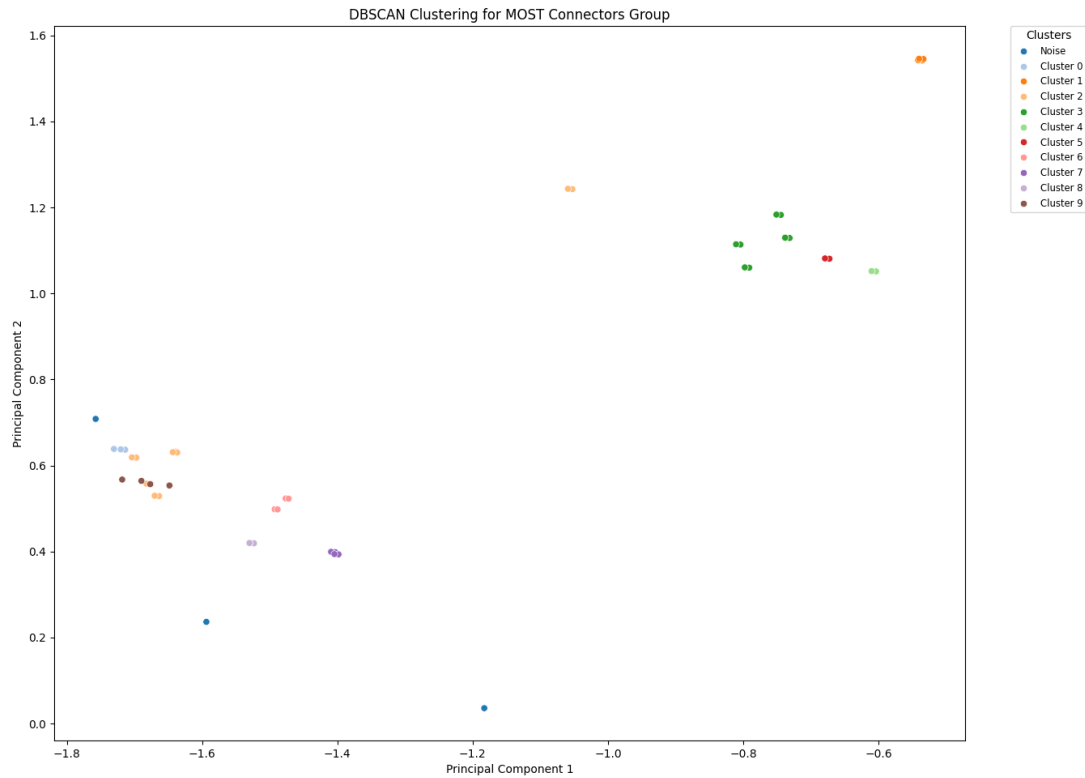


Figure 3.29: Second round of clustering using DBSCAN

It can be seen that the different data points inside the unique clusters in different colors on the PCA plot lie closely together. The algorithm put three data points as outliers based on the chosen parameters.

After cluster 2 was picked up and again, DBSCAN was performed on that cluster to group the connectors once more, as shown in Figure 3.29.

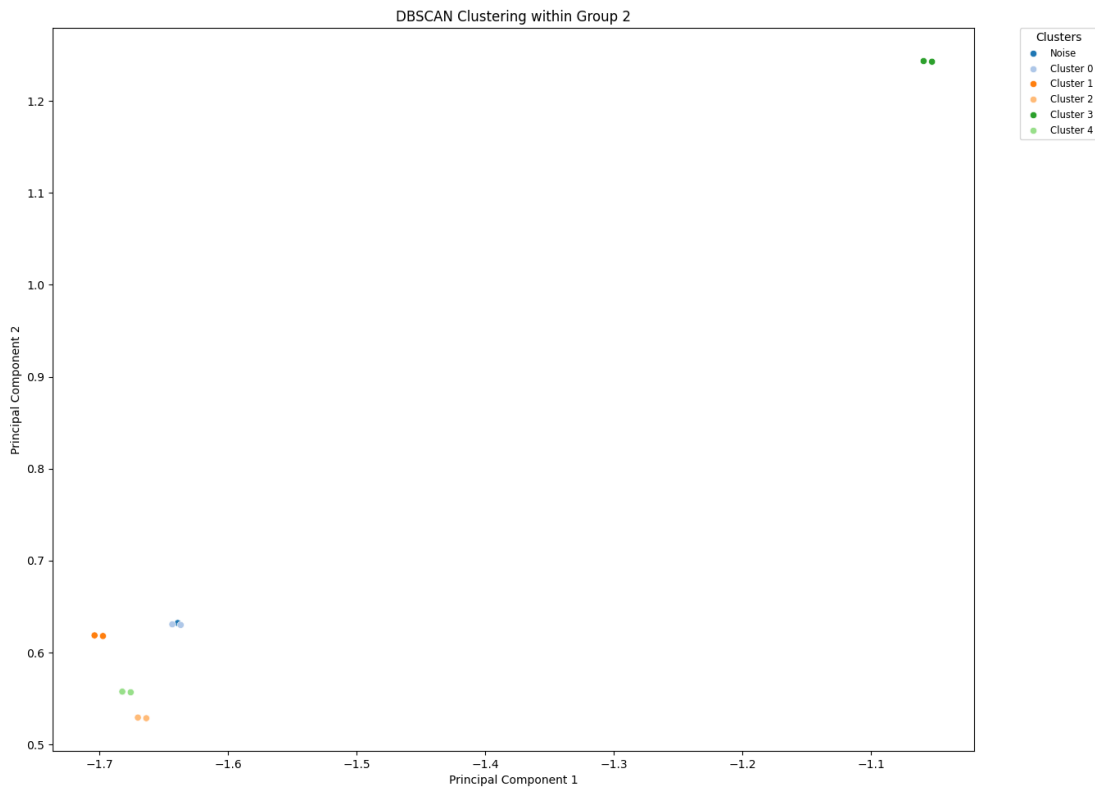


Figure 3.30: Third round of clustering using DBSCAN

Next, for example, the top three closest connectors for each connector in the MOST connector group can be found using the Euclidean distance. This was validated for the MOST connectors using available information from the engineering department at Yazaki. If interchangeable connectors are found, the ones can be chosen with the lowest purchasing price for a larger margin. Lower competitor prices were found for these connectors, saving 53.200 euros when buying a volume of 1.865.280 connectors.

Chapter 4

Conclusion and future work

The first goal of this thesis was to implement and validate machine learning-based clustering techniques to group wire-harness connectors based on their characteristics. Three unsupervised clustering algorithms were implemented on a preprocessed dataset of approximately 6,500 connectors without true labels to group the connectors: K-means, DBSCAN, and GMM. Second, the K-means method was optimized specifically for the connectors by comparing different ways of using that clustering method on a dataset containing 1,020 connectors, including true labels. Finally, the most similar connectors could be found by combining the K-means and DBSCAN clustering methods.

For the dataset of 6,500 connectors without true labels, it was shown that a number of 17 clusters is optimal by calculating the silhouette scores for different clusters. This was confirmed later in this thesis after receiving additional information from the engineering at Yazaki. The results of the clustering techniques were visualized using the PCA and t-SNE dimensionality reduction techniques. T-SNE showed clearer differences between the clusters and was better suited for visualizing the high-dimensional data. Results showed that DBSCAN identifies clusters of different shapes and densities better than K-means. The plots of the results after using GMM demonstrated that the clusters are more compact compared to K-means and DBSCAN. GMM offers a more nuanced understanding of grouping by assigning probability to every data point belonging to a cluster. However, this makes this method computationally more intense. It can also be concluded that performing a decision tree based on the clustering results offers a clear visual interpretation of the decisions made in the clustering process.

The second part of the thesis used a dataset containing 1,020 connectors, including true labels. Five comparisons were made to find the best possible approach for the connectors: with or without standardization, with or without adding weights to the characteristics based on their importance, different initialization techniques, balanced K-means compared to normal K-means, and with or without handling outliers. Based on the results, it can be concluded that using the K-means method, without standardizing the data, without adding weights, using the average values of the connectors per type as the initial center, and handling the outliers gives the best performance of K-means in terms of the confusion matrices and the silhouette scores specifically for grouping the wire-harness connectors. This best approach had a ratio of the correct predictions relative to the incorrect predictions of 117.29%. Also, the silhouette score was 0.203. The accuracy of this method was 62%.

This study's third and last part contained findings on the most similar connectors based on their characteristics. The suggested approach was considered successful after combining the K-means algorithm to group the clusters into 17 groups, selecting one group and performing DBSCAN on that group, and finally selecting one more group after the first DBSCAN grouping and performing DBSCAN for a second time on that group. Finally, the top three closest connectors for each connector in the MOST connector group can be found using the Euclidean distance. This was validated for the MOST connectors using available information from the engineering department at Yazaki.

The most important implication of this master's thesis is finding the most similar connectors to find interchangeable wire harness connectors without going manually through all the databases in Yazaki. If interchangeable connectors are found, the ones with the lowest purchasing price can be chosen so the margin is maximal, and the RfQ response to the OEM can have a competitive advantage in terms of price and time.

Despite this positive implication, the used approaches also have limitations. The main disadvantage of using K-means clustering to group the connectors is that this method assumes that the clusters are of different densities and shapes, which could lead to incorrect results. Additionally, it was stated that removing the outliers improved the K-means method, although, by doing this, some valuable information may be lost. Because additional knowledge of connectors was used to optimize the suggested clustering approaches, the generalizability of the approaches could be decreased.

Future work and research are necessary to develop the suggested approaches and overcome the limitations. This project was performed at an automotive supplier company; thus, in the future, it would be interesting to create a user interface and interactive visualization tools to help the Yazaki employees utilize the clustering techniques effectively. Additionally, the approach used to find the most similar connectors using a dataset of approximately 1,000 connectors could also be used in the future on the dataset without true labels with around 6,500 connectors after preprocessing.

In conclusion, this study underscores the critical role of data-driven business and the use of machine learning in the automotive wire harness business to automate processes like the RfQ process.

References

- [1] Y. Corporation, “YAZAKI Corporation,” YAZAKI Corporation. <https://www.yazaki-group.com/en/> (accessed Mar. 23, 2024).
- [2] L. Chen, “Automotive Wire Harness Connectors: An ultimate guide on wiring connectors for automotive applications,” Wire Harness and Cable Assemblies Manufacturer - Cloom, Sep. 27, 2022. <https://www.wiringo.com/automotive-wire-harness-connectors.html> (accessed Mar. 23, 2024).
- [3] By, “The Surprisingly Manual Process Of Building Automotive Wire Harnesses,” Hackaday, Jul. 27, 2022. <https://hackaday.com/2022/07/27/the-surprisingly-manual-process-of-building-automotivewire-harnesses/> (accessed Mar. 26, 2024)
- [4] R. McAdam and D. McCormack, “Integrating business processes for global alignment and supply chain management,” Business Process Management Journal, vol. 7, no. 2, pp. 113–130, May 2001, doi: 10.1108/14637150110389696
- [5] By, “Key elements of RfQ processes,” Medium, Jul. 11, 2017. https://medium.com/@Hyvan_Technologies/key-elements-of-rfq-processes-d83b3e176e3b/ (accessed Mar. 26, 2024).
- [6] Car - centre for automotive research, “The strategic value of information in the RFQ respons process,” Center for Automotive Research, Aug. 2007.
- [7] “Automotive & Mobility,” Bain, 2019. <https://www.bain.com/industry-expertise/automotive/> (accessed Mar. 26, 2024).
- [8] Amphenol, “The Crucial Role of Connectors in Vehicle Control Units of EVs,” Amphenol communications solutions , Aug. 14, 2023. <https://www.amphenol-cs.com/connect/the-crucial-roleof-connectors-in-vehicle-vontrol-units-ofevs.html#:~:text=Enabling%20Connections%3A%20Connectors%20establish%20secure,functioning%20of%20the%20vehicle's%20systems.> (accessed Mar. 26, 2024).
- [9] “Degradation of road tested automotive connectors,” IEEE Journals & Magazine | IEEE Xplore, Mar. 01, 2000. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=833055> (accessed Mar. 26, 2024).
- [10] “Degradation of road tested automotive connectors,” IEEE Journals & Magazine | IEEE Xplore, Mar. 01, 2000.

- <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=833055> (accessed Mar. 26, 2024).
- [11] “Automotive, Truck, Bus, and Off-Road Vehicle Connectors,” TE Connectivity, Sep. 05, 2019. <https://www.te.com/usa-en/products/connectors/automotive-connectors.html> (accessed Apr. 2, 2024).
- [12] “Automotive and Sealed PCB/Wire Connectors | Molex,” Molex.com, 2024. <https://www.molex.com/en-us/products/automotive-connectivity/automotive-pcb-wire-connectors> (accessed Apr. 4, 2024).
- [13] “Lightening Up: How Less Heavy Vehicles Can Help Cut CO₂ Emissions,” FIA Region I. <https://www.fiaregion1.com/lightcarlowcarbon/> (accessed Apr. 10, 2024).
- [14] TE Connectivity, “CONNECTIVITY IN NEXT GENERATION AUTOMOTIVE E/E ARCHITECTURES,” TE Connectivity. Accessed: Apr. 04, 2024. [Online]. Available: <https://www.te.com/content/dam/tecom/documents/automotive/global/Connectivity-in-Next-Gen-Auto-EE-ArchitecturesWhitepaper.pdf>
- [15] “(PDF) Data Science and Its Relationship to Big Data and Data-Driven Decision Making,” ResearchGate. https://www.researchgate.net/publication/256439081_Data_Science_and_Its_Relationship_to_Big_Data_and_Data-Driven_Decision_Making (accessed Apr. 4, 2024)
- [16] “Data science for business: benefits, challenges - ProQuest,” www.proquest.com. <https://www.proquest.com/docview/2403866220?sourcetype=Scholarly%20Journals> (accessed Apr. 15, 2024).
- [17] Car - centre for automotive research, “The strategic value of information in the RFQ respons process,” Center for Automotive Research, Aug. 2007.
- [18] V. Çetin and O. Yıldız, “A comprehensive review on data preprocessing techniques in data analysis,” *Mu Hendislik Bilimleri Dergisi/Mühendislik Bilimleri Dergisi*, vol. 28, no. 2, pp. 299–312, Jan. 2022, doi: 10.5505/pajes.2021.62687. (accessed Apr. 4, 2024).
- [19] 7.1.6. What are outliers in the data?” <https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm> (accessed Apr. 4, 2024).
- [20] C. A. W. Glas, “Missing data,” in *Elsevier eBooks*, 2010, pp. 283–288. doi: 10.1016/b978-0-08-044894-7.01346-4. (accessed Apr. 4, 2024).
- [21] N. Zhu, “How to select automotive connectors for wire harness?,” Mar. 08, 2018. <https://www.linkedin.com/pulse/how-select-automotive-connectors-wire-harness-nicole-zhu/>. (accessed Apr. 4, 2024).

- [22] B. M. S. Hasan and A. M. Abdulazeez, "A review of Principal Component Analysis Algorithm for Dimensionality Reduction," Apr. 15, 2021. <https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032> (accessed Apr. 2, 2024).
- [23] GeeksforGeeks, "Principal Component Analysis(PCA)," *GeeksforGeeks*, Dec. 06, 2023. <https://www.geeksforgeeks.org/principal-component-analysis-pca/> (accessed Apr. 2, 2024).
- [24] "Introduction to t-SNE," *Datacamp*. <https://www.datacamp.com/tutorial/introduction-t-sne> (accessed Apr. 04, 2024).
- [25] GeeksforGeeks, "Normalization vs Standardization," *GeeksforGeeks*, Nov. 12, 2021. <https://www.geeksforgeeks.org/normalization-vs-standardization/> (accessed Apr. 04, 2024).
- [26] "Using categorical data with one hot encoding," *Kaggle*. <https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding> (accessed May 07, 2024).
- [27] "(PDF) Unsupervised Learning: Clustering," ResearchGate. https://www.researchgate.net/publication/322543367_Unsupervised_Learning_Clustering (accessed Apr. 4, 2024).
- [28] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, Sep. 1999, doi: <https://doi.org/10.1145/331499.331504>. (accessed Apr. 4, 2024).
- [29] GeeksforGeeks, "K means Clustering Introduction," *GeeksforGeeks*, Mar. 11, 2024. <https://www.geeksforgeeks.org/k-means-clustering-introduction/>. (accessed Apr. 4, 2024).
- [30] T. K. m Kodinariya and P. Makwana, "Review on Determining of Cluster in K-means Clustering," Jan. 2013. Accessed: May 05, 2024. [Online]. Available: https://www.researchgate.net/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering
- [31] "2.1. Gaussian mixture models," *Scikit-learn*. <https://scikit-learn.org/stable/modules/mixture.html> (accessed Apr. 4, 2024).
- [32] O. C. Carrasco, "Gaussian mixture model explained," *Built In*, Feb. 23, 2024. <https://builtin.com/articles/gaussian-mixture-model> (accessed Apr. 4, 2024).

- [33] Schubert, E., Sander, J., Ester, M., Kriegel, H.-P., & Xu, X. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Transactions on Database Systems*. (accessed Apr. 4, 2024).
- [34] GeeksforGeeks, "DBSCAN Clustering in ML Density based clustering," *GeeksforGeeks*, May 23, 2023. <https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/>. (accessed Apr. 4, 2024).
- [35] S. Dasgupta, "Understanding the epsilon parameter of DBSCAN clustering algorithm," *Medium*, Mar. 30, 2022. [Online]. Available: <https://medium.com/@saurabh.dasgupta1/understanding-the-epsilon-parameter-of-dbscan-clustering-algorithm-fe85669e0cae> (accessed Apr. 4, 2024).
- [36] Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. *Advances in Knowledge Discovery and Data Mining*. (accessed Apr. 4, 2024).
- [37] M. Mayo, "Centroid Initialization Methods for k-means Clustering - KDnuggets," *KDnuggets*. <https://www.kdnuggets.com/2020/06/centroid-initialization-k-means-clustering.html> (accessed Apr. 4, 2024).
- [38] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. 2009. doi: 10.1007/978-0-387-84858-7. (accessed Apr. 4, 2024).
- [39] N. Mehra, "Decision Tree and Random Forest - Naman mehra - medium," *Medium*, Jan. 06, 2022. [Online]. Available: <https://namanmehra1207.medium.com/decision-tree-and-random-forest-3abd62e49cb5>. (accessed Apr. 4, 2024).
- [40] S. Gorthy, "Euclidean Distance explained," *Built In*, Apr. 01, 2024. <https://builtin.com/articles/euclidean-distance> (accessed Apr. 4, 2024).
- [41] "ML - Analysis of Silhouette score." https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_analysis_of_silhouette_score.htm (accessed Apr. 4, 2024).
- [42] "Steps of the Hungarian algorithm - HungarianAlgorithm.com." <https://www.hungarianalgorithm.com/hungarianalgorithm.php> (accessed Apr. 4, 2024).
- [43] "Selecting the number of clusters with silhouette analysis on KMeans clustering — scikit-learn 0.21.2 documentation," *Scikit-learn.org*, 2019. https://scikitlearn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html (accessed Apr. 4, 2024)

- [44] Z. Jaadi, “When and why to standardize your data,” *Built In*, Aug. 04, 2023. <https://builtin.com/data-science/when-and-why-standardize-your-data#:~:text=Normalization%20involves%20scaling%20data%20values%20in%20a%20range%20between%20%5B0,is%20best%20for%20normal%20distribution>. (accessed Apr. 4, 2024)
- [45] T. Mullin, “DBSCAN Parameter Estimation using Python - Tara Mullin - Medium,” *Medium*, Dec. 15, 2021. [Online]. Available: <https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd>. (accessed Apr. 4, 2024)