UNIVERSITY OF TWENTE.

**Faculty of Electrical Engineering,
Mathematics & Computer Science**

Automated Electronic Component Selection: A Machine Learning approach

Mohd. Abbas Rizvi
Final Thesis
MSc. Business Information Technology

Graduation Committee:
prof. dr. F.A. Bukhsh
prof. dr. A. Abhishta
prof. N. Bouali

Faculty of Electrical Engineering,
Mathematics & Computer Science
University of Twente
Drienerlolaan 5
7522 NB Enschede
The Netherlands

Abstract

The research work in the study is conducted at Signify N.V., a multinational Lighting corporation providing lighting products and solutions to the consumer all over the globe. Signify is seeking ways to improve electronic component recommendation process, with the aim to reduce process complexity and product delivery time. The aim of this research is to automate the component selection process by developing a model to predict the best electronic component to be used in the design by designer/engineer based on their technical parameter requirement. The model should recommend the optimal component considering technical as well as commercial aspect of the component. A Hybrid supervised-unsupervised model approach is investigated for predicting the best electronic component. The Cross Industry Standard Process for Data Mining (CRISP-DM) is used across the study.

Clustering techniques such as Hierarchical agglomerative, BIRCH, DBSCAN, OPTICS and Mean Shift are used to group similar components. After clustering, supervised learning algorithms such as Support vector machine, K-nearest neighbor, Random Forest, XGBoost and Naïve bayes are applied to predict the optimal component.

Hierarchical agglomerative clustering and K-Nearest neighbor had the best result compared to the rest methods and hence were selected for model development. A hybrid model using combination of agglomerative clustering and KNN approach was developed. The predicted component results were evaluated and compared with the existing method. The model predicted 84% accurately for capacitor dataset and 81% accurately for resistor dataset. The model also predict substitute for obsolete component which would prevent the inclusion of obsolete component in the design. The model was also deployed on Heroku platform to complete the CRISP-DM methodology cycle.

The proposed model would help component engineers and designer by saving a lot of their time which is needed for approving the design using traditional manual method. In this way, the overall product development or delivery time can also be reduced.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Signify is a company which provides Lighting solutions to customers and consumers globally. It was formerly known as Philips Lighting. Signify is the world's largest manufacturer of lighting for professionals, consumers, and the Internet of Things. Customers can experience a higher quality of light owing to the energy efficient lighting products, systems, and services, which make people's lives better and far more comfortable, businesses more creative, and cities more sustainable. The Component Engineering department of the company deals with the decision-making process involving the selection of best suitable components for the end electronic product considering commercial and technical parameters.

Electronic components are an important aspect of the design of electronic systems, and they are continually evolving in terms of product quality and size. Successful organizations and products, without a doubt, have used the most optimal components in the designs [1]. This results in a need for the technical engineering team to gain the knowledge to choose the best optimal electronic component. Although it looks that selecting components for hardware design needs is a basic and uncomplicated process, there are numerous considerations and "techniques" involved [1]. A mistake in component selection might result in catastrophic consequences, including rejection and dismissal of the overall project and product. As a result, extreme caution and safeguards are recommended when engaging in this action. As a result, design engineering department is helped by component engineers in selecting the suitable component for their respective designs. Component engineers must have a broad awareness of components, their specifications, architectures, and functionalities, as well as extensive knowledge and expertise in these areas [1].

Abundance amount of data availability has led to the upgradation of existing old traditional component selection process in supply chain domain. Due to large amount of supplier willing to supply diverse components to the firm, a need for more advanced and intelligent component selection process has arisen which can be found effective

in predicting or recommending the best electronic component [2]. Machine learning is a method used for predictive modelling. An approach to build a AI model using machine learning could prove to be helpful for the electrical designers and component engineer. The developed model will predict the most optimal component to be used for a circuit design which will lead to a flawless designing process.

To predict the best component, an appropriate strategy is vital. The first step to predict the optimal electronic component is to cluster the similar components as per their technical parameters. The availability of data makes this prediction possible. Examples of data sources are selling data, procurement data, supplier data and Customer Relationship Management system data. Next, clustered similar components can be ranked as per commercial business rules such as low price, low lead time and high commonality. Commonality refers to the number of various electronic designs in which selected components can be used which will help companies to procure in huge quantities and at a low-price rate from suppliers. This setup enables the firm to focus on optimal components being used in the design and manufacturing of PCBs and also reducing the final product delivery time of the product/project. In this study, the focus is to predict the best electronic component to be used/selected by the electrical designer in their circuit design considering technical and commercial aspects of the business which will help in enhancing the company revenue and profit.

To effectively predict which electronic components to be used, data mining techniques can be applied. Data mining methods are classified into two categories viz, supervised and unsupervised learning. In the unsupervised learning problem, only the features X are observed and there is no measurement of the outcome y. The goal is to identify the data's grouping or clustering. The attributes X and the outcome y for a group of objects are observed in the supervised learning problem. The objective is to accurately predict the outcome y for new unknown items as feasible as possible [5].

Unsupervised and supervised machine learning approaches are integrated in hybrid data mining models, which has been shown to improve prediction results [6]. In this work, component prediction will therefore be performed in two learning stages. In the first stage, homogeneous groups of electronic components will be identified using unsupervised learning. In the second stage, using supervised learning, a classifier will be developed for each cluster to predict which components are optimal. The

expectation is that the hybrid approach will improve the predictive performance by combining unsupervised and supervised learning.

1.1 Problem Statement

The highly competitive and dynamic semiconductor market results in enhancing the traditional product development techniques of companies. The choice of the correct component from the supplier plays a role in determining performance and consistency of a device. A well-designed process to predict the best optimal electronic component to be used in design may aid in the shortlisting of a promising supplier, and a collaborative effort by various cross-functional is vital to choose an optimal component. Electrical, mechanical, environmental, cost, dependability factor, component life cycle, and other factors should all be considered at the very least when selecting a component [3].

The team of component engineer’s is responsible for optimal component selection because they function as a connection between the electrical design engineer, material procurement and product assembly/manufacturing units [3]. The most crucial function in the product development cycle is selecting the appropriate component because a poor choice here results in significant and unanticipated losses (money, machine, personnel, time, etc.) throughout the development cycle, resulting in component failures and delays.

Another factor to consider when choosing a component is the component's life cycle, which is especially important in the case of electronic circuits [1]. In addition to the abovementioned characteristics, the expected life of any of the parts used inside the final developed product defines the product's life [3]. Choosing a part that is approaching end-of-life is a concern or an expenditure for the organization and choosing a part that was recently produced puts the product's sustainability at danger. As a result, sound judgment should be used in the BoMs (bill of materials) to encourage the use of long-life components. To avoid the product's unexpected threats, maintaining a track of alternate parts and resources must be appreciated.

The component's "reliability," defined as the possibility of the component executing its anticipated task over the defined time period is the next key factor to evaluate [1]. To

put it another way, a component's reliability is the consistency of its functioning across time. The reliability parameter is also used to compute guarantee and safety durations, as well as make design alternative decisions.

1.2 Research Questions

The general research goal to achieve in this thesis is:

Main goal: Evaluate the value of combining unsupervised and supervised machine learning to predict the optimal electronic component.

To that aim, the following sub-research Questions have been developed to help achieve the main research goal of predicting the best optimal electronic component.

Research Question 1: Which of the existing machine learning approaches could be used to predict the best electronic component?

Research Question 2: To predict the best electronic component using machine learning method, which algorithms could prove to be the most efficient?

Research Question 3: Can the proposed automated component selection approach using Machine learning method perform better compared to the traditional manual component selection approach?

Research Question 4: Can the developed ML model be able to predict accurate substitutes for obsolete electronics components?

1.3 Contribution

1. In this thesis, a machine learning model using a combination of unsupervised and supervised learning ML techniques such as clustering and KNN, SVM, Random Forest, Naive bayes, XGBoost is developed to predict the best electronic component such as capacitor, resistor to be used in the circuit design. According to our knowledge, this research is a novel approach that puts the aforementioned Machine Learning (ML) methods to work in the semiconductor industry's component engineering domain.

2. Also, the research thesis proposes and evaluates a number of machine learning (ML) algorithms for dealing with data quality issue. Normalization of the raw data using data transformation method and dealing with missing data is discussed.

3. This study helps to analyze and compare how machine learning models can be used to replace the traditional human judgment method in component engineering. Performance of ML model in reducing complexity of the existing process and reduction in product/project delivery time challenges will be discussed. The approval of a design to be done by a component engineer takes on average 5-6 hours which can be resolved using the model in a few seconds. This will help in reducing delivery time and also reducing the 'approval process' complexity as now the designer would input the component suggested/recommended by the model itself whose results are validated by the component engineering team.

4. To our knowledge, this is the first attempt to automate the component selection process using Machine learning approach considering both technical factors such as capacitance value, resistance value, voltage, tolerance, etc. and commercial factors such as low price, low lead time and high commonality with the aim to predict the best components to be used in design. Few studies have shown component recognition/selection using image recognition Deep learning methods on image dataset, but none have provided a model dealing with dataset with above technical and commercial factors.

5. The hybrid model developed using unsupervised and supervised ML techniques will be helpful for companies to identify which supplier components are frequently used in their circuit design. This data can be forwarded to the negotiation/pricing team of the company which can be proved to be vital in negotiating with suppliers for low prices in the next procurement cycle.

1.4 Outline

The report structure is as follows: Discussion on the existing use of hybrid machine learning models using the results from systematic literature review is performed in chapter 2. Eventually in chapter 3, we enlist a more detailed discussion related to data preparation. In chapter 4, a methodology for this study is presented. Chapter 5

emphasizes on the various modelling techniques used. Thereafter, in the next chapter 6 the developed model results are discussed and their performance is evaluated. Chapter 7 present the conclusion of this study, also the limitations and recommendations for this project are discussed in this section.

Chapter 2

Literature Review

This section presents a description of the electronic component selection process and the related work performed in relation to the thesis subject. Section 2.1 presents a overview definition and explanation of the electronic component selection process respectively. Thereafter, in section 2.2, description of data mining techniques is discussed. In section 2.3 the background work in relation to component and supplier selection is illustrated. Conclusion of this chapter discusses about the overall summary of the literature review performed in section 2.4.

2.1 Electronic component selection process

Component selection is the method of choosing an appropriate component or a group of similar components from several providers to enable the built electrical circuit to accomplish its desired purpose. The component engineer may need to first comprehend the circuit features and functions or obtain the exact parametric values. To expedite component selection, the component engineer must strive to acquire as many information as necessary from the design engineer. The design engineer almost always offers only the set of important specifications, but the component engineer must additionally consider the non-listed specifications. As a result, engaging with the electrical design engineer team is crucial for the component engineer, whenever he or she has a doubt, rather than guessing or taking parameters for assumption [3]. Because he functions as a connection between procurement chain, design, and manufacturing teams, and enhances the supply chain operations, the component engineer is preferred for component and supplier selection. The most crucial duty in the product development cycle is selecting the proper component, because a poor choice here leads to significant and unanticipated losses (money, personnel, time, etc.) throughout the building process, resulting in product failures and delays.

The component engineering team at Signify assists the electrical designer/engineer in selecting or recommending appropriate components for use in electronic circuit designs. The electrical designer submits his or her design for approval to the component department. A component engineer is responsible for understanding the properties of each component utilized in the circuit design and recommending a replacement component based on technical and commercial considerations. If any component in the design has to be replaced, the information is provided to the designer, along with a design modification to be made. The design is authorized by the component engineering department and forwarded to the procurement team after the necessary amendments are made. Hence, the component engineer is the one who deals with product design and supplier selection activities. The commercial parameters to be considered while selection are Low price, Low generic envelope and high commonality. Electrical, mechanical, and environmental factors are the three different types of factors which influence the performance of electronic component [1].

Component operating voltages, capacitance value, operating temperature, memory width, cut-off frequency are some examples of component electrical parameters. These variables are those that strongly influence the component's performance. Each component category is evaluated and designed for its own primary criteria.

Mechanical parameters are those that describe the component's physical shape. In nature, mechanical factor variables could be sensed [3]. The characteristics of the component determine how the component interacts with other circuit board components. Because almost all vendors have the same functional component in multiple package forms, it becomes quite significant to possess information about specifications of the component from the designer in the early stage of the project [1].

2.2 Data Mining

Data mining is an important methodology used for extracting meaningful insights or hypotheses from large amounts of data. It is a process of retrieving usable insights from the complex raw dataset by focusing on data analysis techniques. Data mining involves processes such as understanding and preparation of data, reducing the data

or transforming into usable format and lastly extracting useful insights from the data. [4] has proposed following basic steps involved in Data Mining:

1. Identifying the goal from customer's perspective.
2. Extraction of target data to be used from the database.
3. Building strategies and techniques to deal with missing values in the dataset.
4. Reduction of dataset features by extracting the relevant features/variables using dimensionality reduction or different transformation techniques.
5. Processing the cleaned and transformed data into relevant machine learning models to be used.

Data mining consists of various techniques such as classification, prediction, clustering, pattern recognition. It is the process of extracting patterns, correlations, from data. Because the human brain is unable of processing large amounts of data efficiently and correctly, data mining technologies save effort and decrease the danger of human mistake. Machine Learning plays an important role in data mining process. The insights extracted from the data using data mining process are used to fit several machine learning models based on industry use case. In the semiconductor sector, data mining techniques have a wide range of uses. Quality control, supplier selection, production, decision making systems, and instrumentation are primary areas for data mining applications in semiconductor manufacturing [5].

2.2.1. Machine Learning

Machine learning is a group of algorithms that retrieve features from data. In exploratory data analysis, a technique like this is a complement to the models. The learning methodologies of Machine Learning based solutions may be divided into three categories: supervised learning, unsupervised learning, and reinforcement learning. The labeling of the incoming data distinguishes supervised from unsupervised learning. Models developed using supervised or unsupervised learning do not interact with their surroundings whereas models developed using reinforcement learning technique interact with their surroundings by taking actions and receiving rewards, with an objective of increasing the overall reward. All of these methodologies are applicable to Exploratory data analysis challenges.

Machine Learning has three basic functions: descriptive (where it analyzes data to offer insights), predictive (where it anticipates the future outcome), and prescriptive (where it forecasts future outcome and also makes recommendations for actions based on the data).

2.2.2. Unsupervised Learning

Clustering is the unsupervised machine learning technique of recognizing natural groups or clusters within multivariate data based on certain similarity metric such as Manhattan distance, Euclidean distance [6]. A cluster center which is known as centroid is commonly used to determine a cluster. Clustering of data points is a challenge in unsupervised machine learning because data clusters can be of various forms and sizes [6]. Applications of clustering techniques are in the field of image segmentation [7], machine learning and data mining. Unsupervised machine learning method such as clustering are divided into two categories: hierarchical clustering algorithms and partitional clustering algorithms [8]. Hierarchical clustering techniques use systematic splitting or merging approaches to create a cluster tree known as dendrogram [9]. Hierarchical clustering algorithms follows two different approaches viz. divisive and merging. In divisive method, a top-down approach is followed, wherein all datapoints are clustered together in one cluster at first and then splits are performed significantly while moving from top to bottom. The merging method is also known as agglomerative clustering in which bottom-up approach is used. In this method, each datapoint is treated as a single group at first, and subsequently identical datapoints are grouped into a single cluster until only one cluster remains. Divisive techniques begin with a single cluster of all sample points and split the most optimal cluster at each iteration until a stopping condition of desired number of clusters is met. In the case of agglomerative algorithms, on the other hand, each item begins as a single cluster, and the most identical pair of clusters is merged at each step. According to the manner they determine cluster similarity, agglomerative hierarchical clustering algorithms are classified as single-link, complete-link, or average-link. The clusters with the least distance between their nearest patterns are merged using single link techniques. On the other hand, complete link algorithms combine clusters with the least distance between their most faraway patterns [10]. Partitional clustering techniques partition the data into a collection of clusters. These methods strive to

minimize a square error function and hence can be termed as optimization methods [11]. Partitional clustering techniques cluster the data into partitions that are generally optimized locally. Popular partitional clustering algorithms used are k-means and its variant bisecting k-means.

2.2.3 Supervised Learning

The most widely used type of Machine learning technique is Supervised Learning [12]. Supervised learning is a type of machine learning technique wherein the systems are taught with labeled data and after getting trained with the labelled dataset, output is determined using that data. Some of the input data has already been labeled with the proper output. This learning method seeks to identify a relationship between the input and output variables to find a pattern and then use this pattern knowledge to predict the output for an unseen test dataset. Two majorly critical tasks in supervised machine learning are classification and regression [13].

A number of machine learning algorithms have been created during the last decade, in tandem with the expanding amount of data accessible for research. Depending on whether or not they include labelled data as target variable, they might be classified as supervised or unsupervised learning technique. If the dataset has a labelled output data using which the algorithm can learn and get trained, then these approaches are referred to be supervised; otherwise, they are referred to as unsupervised or clustering techniques [14]. The supervised learning techniques can be divided in to six groups as mentioned below, Artificial neural network, KNN, Decision Tree, Random Forest, SVM and Naive Bayes.

Table 1. Supervised learning techniques

Algorithm Type	Example
Logic based algorithms	C4.5
Perceptron-based techniques	Artificial Neural Network
Statistical learning algorithms	Naïve Bayes classifiers
Instance-based learning	K-Nearest Neighbor algorithm
Support Vector Machines	Support Vector Machines
Regressions	Logistic regression

Table 2. Comparison of supervised machine learning techniques

	Decision Trees	Neural Networks	Naive Bayes	kNN	SVM
Accuracy in general	**	***	*	**	****
Speed of learning with respect to number of attributes and the number of instances	***	*	****	****	*
Speed of classification	****	****	****	*	****
Tolerance to missing values	***	*	****	*	**
Tolerance to irrelevant attributes	***	*	**	**	****
Tolerance to redundant attributes	**	**	*	**	***
Tolerance to highly interdependent attributes (e.g. purity problems)	**	***	*	*	***
Dealing with discrete/binary/continuous attributes	****	***(not discrete)	***(not continuous)	***(not directly discrete)	**(not discrete)
Tolerance to noise	**	**	***	*	**
Dealing with danger of overfitting	**	*	***	***	**
Attempts for incremental learning	**	***	****	****	**
Explanation ability/transparency of knowledge/classifications	****	*	****	**	*
Model parameter handling	***	*	****	***	*

2.3 Application of machine learning to predict the best electronic component in the semiconductor industry

In this section we highlight and study the work related to prediction/recommendation of the best electronic component in the semiconductor domain using Machine Learning techniques.

2.3.1 Methodology

A comprehensive literature review based upon the work and criteria of [15] was conducted with the goal of exploring similar work on this issue. The research approach for finding, extracting, and evaluating relevant papers is described in the following paragraphs. Only database available through the University of Twente were taken into account. Scopus and Web of Science were chosen as the primary sources.

Search Query

A list of keywords relating to the study topics are used to create the search query. The relevancy of the main research question, as well as the sub-questions, is used to determine the major keywords. The use of hybrid, for example, is intended to help answer the study's last sub-question on the technique used.

Scopus Query:

TITLE-ABS-KEY ((”Machine Learning”) AND (”Techniques” OR ”Approach” OR ”Methods”) AND (”model” OR ”models”) AND (”predicting” OR ”prediction”) AND (”ranking”) AND (”hybrid”) AND (”supervised”) AND (”unsupervised”) AND (”ensemble”)) AND (LIMIT-TO (PUBSTAGE , ”final”)) AND (LIMIT-TO (DOCTYPE , ”ar”) OR LIMIT-TO (DOCTYPE , ”re”)) AND (LIMIT- TO (SUBJAREA , ”COMP”) OR LIMIT-TO (SUBJAREA , ”ENGI”)) AND (LIMIT-TO (PUBYEAR , 2021) OR LIMIT-TO (PUBYEAR , 2020) OR LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018)) AND (LIMIT- TO (LANGUAGE , ”English”))

Web of Science Query:

TS=((”Machine Learning”) AND (”Techniques” OR ”Approach” OR ”Methods”) AND (”model” OR ”models”) AND (”predicting” OR ”prediction”) AND (”ranking”) AND (”hybrid”) AND (”supervised”) AND (”unsupervised”) AND (”ensemble”))

The search was carried out by looking up the phrases in the title, abstract, and keywords. The following constraints were used to define the study's scope.

1. Only articles published after 2018 were deemed suitable, as earlier works, particularly in the fields of science and technology, eventually become outdated.

2. Articles from Computer Science and Engineering subject area domain are considered.

3. Articles only in English language were selected. After defining the above constraints, 262 papers were extracted.

Selection Process

Inclusion and exclusion criteria are defined to help narrow down the results and identify the most important and appropriate research studies. They are as follows:

1. The topic of research area should be focused on hybrid machine learning approach.

2. The studies are based on the approach which illustrates the use of combination of supervised-unsupervised machine learning techniques.

3. The studies complete text is accessible to read online.

4. Articles with the same title or content are not included.

The title, abstract, and keywords of all publications were reviewed using the set of criteria. As a consequence, 11 papers met the selection criterion, 41 had objections, and 210 were deleted because they did not match the requirements. A careful evaluation of the 41 articles in question was conducted by evaluating each portion of the paper, and 16 studies were ultimately selected. Systematic Literature review yields a total of 27 papers.

An overview of the Literature review process is mentioned below:

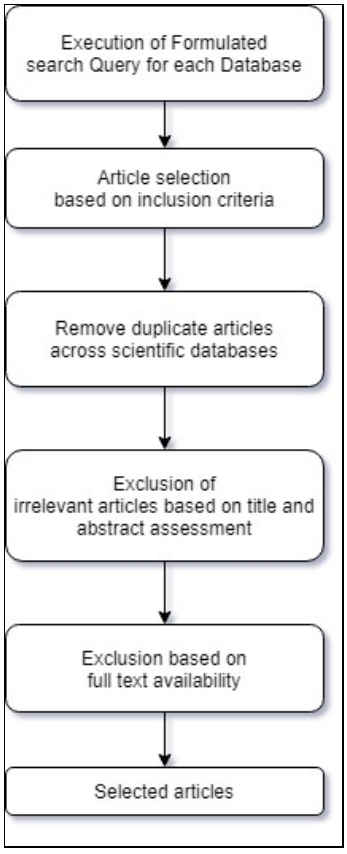

Figure 1. Systematic Literature review process

2.3.2 Previous background work

Following the article selection, important information that is necessary for answering the research question should be gathered. This extracted information will also assist in developing the reference model that will be covered in detail in the future chapter. The following Table 3. lists the 27 papers that were chosen, along with their research goal, research technique, and study outcome. A wide range of hybrid models approach work is carried out in last decade. Combining clustering with a decision tree algorithm is the most straightforward approach for creating a hybrid model. For example, [16] suggested a hybrid model by combining K-means+ID3 algorithm. This approach uses K-means clustering to divide the training data into K distinct clusters, after which each cluster is trained with an ID3 decision tree. Finally, the below mentioned two principles are enforced to each test sample to arrive at a final categorization decision: 1) the nearest neighbor principle, and 2) nearest consensus principle. It was revealed through experiment results that the hybrid model outperforms a conventional ID3 model.

To forecast customer churn rate, [17] suggested a two-stage hybrid model (KM-Boosted C5.0) comprised of an unsupervised clustering algorithm and a boosted C5.0 decision tree. The samples are clustered using a clustering method during first step, after which the segmented labels are included in dataset as a new feature. The newly collected dataset is utilized to train the boosted C5.0 decision tree model in the second step. In comparison to a standard example in which no clustering data is used, their experiment results show that including clustering data improves top decile lift for the hybrid model.

[18] presents a tree bagging and weighted clustering hybrid approach that combines a decision tree with a clustering technique. First, decision tree bagging is used to choose key characteristics and their weights; the weighted attributes are then utilized to form clusters from which future objects are categorized. The TBWC model obtains a greater accuracy than the C4.5 decision tree, according to the findings of the study, which were based on five different datasets.

Furthermore, several research have illustrated both unsupervised and supervised learning to improve neural network performance. [19] presented FC-ANN, a hybrid

intrusion detection method based on a back propagation neural network (BPNN) and fuzzy c-means. Fuzzy c-means clustering is used to partition the training set into different subgroups. Separate BPNN models are then trained as different reference models depending on multiple training groups. Finally, the results are consolidated using a meta learner, specifically a fuzzy aggregation mechanism. KDD Cup 1999 dataset is used in this study. The suggested methodology outperforms a BPNN and algorithms, such as Naive Bayes and Decision tree.

[20] presented the spectral clustering and deep neural network (SCDNN) technique, which integrates spectral clustering and a deep neural network. This approach resembles a hybrid approach of combining clustering and deep neural network. The research area in this study is focused on intrusion detection datasets. In this, firstly the training dataset is partitioned into K subsets to find the cluster center of each group. Next, Deep neural network models trained using the subset. Finally, the test dataset is partitioned into K subgroups using the K cluster centers, and each subset is put into the most suitable Deep neural network model to evaluate the hybrid approach's accuracy. On evaluation, it is observed that the detection accuracy of spectral clustering and deep neural network (SCDNN) model outperformed the Back propagation neural network, Support vector machine and Random Forest.

[21] develops a hybrid hierarchical artificial neural network model for isolating the faults of the Tennessee–Eastman process (TEP). In this approach, fault pattern space is grouped into a few sections using fuzzy c-means clustering. After that, a particular multilayer perceptron (MLP) neural network is developed for each portion to detect the subsection flaws. Finally, to identify which particular Multilayer perceptron neural network should be activated, a supervisor Multilayer perceptron neural network is created. Analyses using simulation datasets revealed that the suggested hybrid technique outperforms a single Multilayer perceptron neural network significantly.

[22] suggested a hierarchical clustering-based approach to tackle the supplier selection problem. In this study a novel distance formula is proposed. The proposed distance formula is compared with existing distance measure formula used in the clustering algorithm such as Euclidean, Manhattan etc. Using the proposed distance formula, a novel hierarchical clustering-based approach is used to recommend the

best supplier. The final results show the proposed hierarchical clustering using new proposed distance formula performs better than interval type-2 fuzzy sets.

[23] proposed a hybrid approach to identify churn credit card holder customer. In this study, a model is developed using clustering and supervised classification techniques. Rough k-means clustering is used to combine with support vector machine algorithm to develop a hybrid model which outperforms other techniques.

Table 3. Literature Findings

No.	Source	Title	ML Techniques
1	[24]	“This paper proposed a hybrid data-level method to handle multiple data impurities like class imbalance, noise and different data distributions within classes. The proposed approach works in phases; in the first phase, it identifies and removes noise from the data, and then, it detects minority and majority clusters by using a kernel-based fuzzy clustering approach.”	Kernel-based fuzzy clustering approach, kernel-based fuzzy clustering approach.
2	[25]	"This paper proposes a hybrid model using unsupervised clustering for prediction of customer churn.”	K-means clustering and Decision Tree.
3	[18]	“This study proposes a hybrid classification framework based on clustering (HCFC). First, it applies a clustering algorithm to partition the training samples into K clusters. Second, it constructs a clustering-based attribute selection measure, i.e., the hybrid information gain ratio, and then trains a C4.5 decision tree based on the hybrid information gain ratio.”	Hybrid classification framework based on clustering (HCFC) and Decision Tree.
4	[26]	“This paper proposed a hierarchical clustering-based method to solve a supplier selection problem and find the proximity of the suppliers.” “This paper presents a Novel Electronic Component Classification Algorithm Based on Hierarchical Convolutional Neural Network.”	Hierarchical convolutional neural network (NH-CNN)
5	[27]	Presentation of a Recommender System with Ensemble Learning and Graph Embedding: A Case on Movie Lens “In this research, group classification and the ensemble learning technique were used for increasing prediction accuracy in recommender systems.”	Decision Tree and ensemble learning techniques were used for increasing prediction accuracy in recommender systems.
6	[16]	“In this paper, we present "k-means+ID3", a method to cascade k-means clustering and the ID3 decision tree learning methods for classifying anomalous and normal activities in a computer network, an active electronic circuit, and a mechanical mass-beam system.”	K-means clustering, ID3 decision tree, K-nearest neighbors.
7	[28]	“This paper presents a new classification algorithm which is a combination of decision tree learning and clustering	Combination of decision tree learning

		called Tree Bagging and Weighted Clustering (TBWC). The TBWC algorithm was developed to enhance the classification performance of a clustering algorithm."	and clustering called Tree Bagging and Weighted Clustering (TBWC).
8	[17]	"In this paper, we use two-stage hybrid models consisting of unsupervised clustering techniques and decision trees with boosting on two different data sets and evaluating the models in terms of top decile lift."	K-means, Hierarchical clustering, Decision Trees.
9	[29]	"This paper presents a hybrid technique, combining simulation and machine learning and examines its applications to data-driven decision-making support in resilient supplier selection."	k-Nearest Neighbors, Logistic Regression.
10	[20]	"This paper proposes a novel approach called SCDNN, which combines spectral clustering (SC) and deep neural network (DNN) algorithms. First, the dataset is divided into k subsets based on sample similarity using cluster centers, as in SC. Next, the distance between data points in a testing set and the training set is measured based on similarity features and is fed into the deep neural network algorithm for intrusion detection."	Spectral clustering, Random forest, Support Vector Machine, Deep Neural Network.
11	[30]	"This paper presents a customer segmentation using clustering and data mining techniques."	K-means clustering, K-nearest neighbors.
12	[31]	"A data-driven approach for analyzing semiconductor manufacturing big data for low yield diagnosis purposes for detecting process root causes for yield improvement."	Random Forest.
13	[32]	"Machine Learning for Predictive Maintenance: A Multiple Classifier Approach."	Support Vector Machines, k-Nearest Neighbors-means Clustering.
14	[33]	Natural Hierarchical Cluster Analysis by Nearest Neighbors with Near-Linear Time Complexity	Agglomerative hierarchical cluster algorithm
15	[34]	"Combining Clustering with Classification: A Technique to Improve Classification Accuracy."	Hierarchical clustering-means clustering, Naive bayes classifier.
16	[35]	"This paper presents Hybrid data mining models for predicting customer churn using clustering and ANN."	Hierarchical clustering-means clustering, Artificial Neural network.
17	[36]	"Machine Learning for Electronic Design Automation."	XGBoost, SVM, Random Forest.

18	[19]	“A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering. In this paper, we propose a new approach, called FC-ANN, based on ANN and fuzzy clustering.”	Back propagation neural network (BPNN) and fuzzy c-means clustering.
19	[37]	“Improving Classification Accuracy Using Clustering Technique.”	K-means clustering, K-nearest neighbor.
20	[38]	“This paper presents a hybrid model using data envelopment analysis (DEA), decision trees (DT) and neural networks (NNs) to assess supplier performance.”	Decision Tree, Deep Neural network.
21	[39]	“This paper presents a review of related work on Machine Learning in Semiconductor Manufacturing and Assembly Lines.”	Clustering, Deep Neural network, SVM, Random Forest.
22	[40]	“K-Means Clustering-Based Electrical Equipment Identification for Smart Building Application. In this paper, we propose a k-means clustering-based electrical equipment identification toward smart building application that can automatically identify the unknown equipment connected to BIoT systems.”	K-means clustering
23	[41]	“This paper describes our work using Artificial Intelligence (AI) techniques to classify components based on an ideal component specification.”	C4.5 Decision tree, Artificial neural network.
24	[42]	“An Electronic Component Recognition algorithm based on Deep Learning with a Faster Squeeze Net.”	Convolutional Neural Network.
25	[43]	“Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. The proposed strategy combines SOM and seven supervised learning methods.”	Self-Organizing map and supervised machine learning.
26	[23]	“This paper presents the combination of combined rough K-means clustering with five supervised classification models, to construct different versions of hybrid models.”	K-means, Decision tree, Support vector machine, Random forest, Naive Bayes-nearest neighbor.
27	[21]	“Designing a hierarchical neural network based on fuzzy clustering for fault diagnosis of the Tennessee–Eastman process”	Fuzzy c-means clustering algorithm and MLP Neural Network.

2.3.3 Summary

Although many studies discussed the topic of prediction using hybrid supervised-unsupervised machine learning approaches, not much has been explored in the component engineering domain for semiconductor industry. The feasibility of using Machine learning techniques in automating the electronic component selection process are yet to be thoroughly explored. Most of the studies conducted have focused on electronic component recognition using Deep Neural network on Image dataset [44]. According to the findings of this study's review, there hasn't been much research on the application of machine learning techniques to the component engineering domain in the semiconductor sector. Hybrid model learning offer a solution to this problem. The majority of the study has been conducted in the supplier selection process in supply chain management, whereas the most essential portion of the component selection process is still done using traditional manual techniques.

The function and necessity of optimal electrical component selection/recommendation are described first in this study, and then several machine learning algorithms applied for various disciplines are retrieved using chosen literature findings. Following that, we looked at several supervised-unsupervised learning combination models, their effectiveness and their application in various fields. Following the extraction of an appropriate Machine Learning Technique from the literature, a supervised-unsupervised combination is employed to construct a Machine Learning model to predict the best component.

Chapter 3

Research Methodology

Data mining is defined as the process of transforming raw data into usable insights [45]. The CRISP-DM (Cross Industry Standard Process for Data Mining) is used to execute the data mining projects.

3.1 CRISP-DM Process Model

A successful data science project should have a dependable and consistent procedure that anybody with a basic understanding of data science can follow and grasp. CRISP DM Methodology can be used as a template to guarantee that all of the distinct factors vital to the data science project is addressed. The data mining method is divided into a limited number of steps at the top level known as phases. Each phase is made up of multiple typical generic tasks. The next level is where you explain how generic actions should be performed in specific conditions. At the end, process instance step comprises of the activities, choices and outcomes of a data mining project.

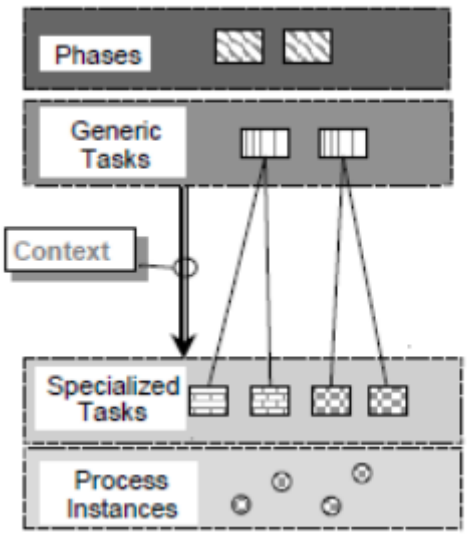

Figure 2. Four stages of CRISP-DM methodology [46]

3.1.1 CRISP-DM

In the data mining sector, CRISP-DM is the most extensively utilized data analysis model technique. The CRISP-DM model includes gives a summary of all different operations involved in a data mining project [46]. This methodology consists of six different steps as shown in Fig 3.

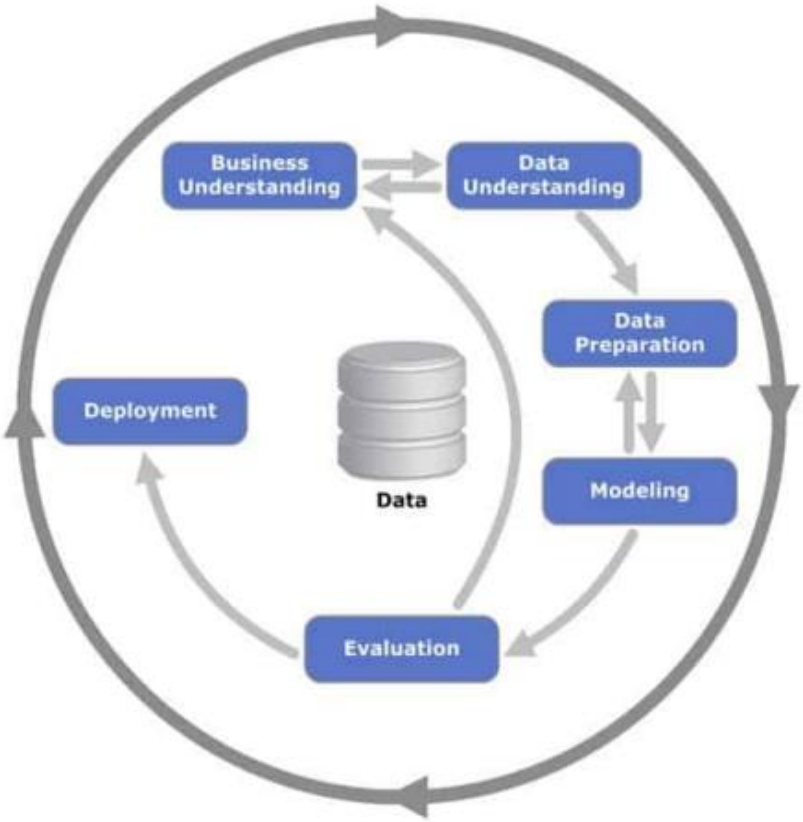

Figure 3. Phases of CRISP-DM methodology [46]

Business understanding:

The business understanding will focus on understanding the business goal for the study. From signify point of view, it is important to understand why for signify is it important to automate the electronic component selection process using machine learning. A literature review study is performed to identify the machine learning techniques which could prove relevant to the business problem.

Data understanding:

Once a proper understanding of business is performed, exploratory data analysis is performed to understand the electronic component data set provided by Signify. Problems related to quality of data such as invalid data format, non-numeric data in numeric features column, missing values are dealt in this section.

Data preparation:

Data preparation is a process wherein data is prepared and made ready to be fed in the model. In this step, new features are made, data is scaled or normalized, and cleaning of dataset is done to get the prepared dataset to be used in Modeling.

Modeling:

In this phase, different machine learning techniques are finalized to be used in respect to the business problem. In this study, a hybrid approach of using clustering technique with supervised learning technique is used. Based on the best combination, a specific hybrid model will be developed.

Evaluation:

In this stage, results of different model combination developed in the abovementioned step is evaluated. Models are evaluated based on the results of key metric parameters and the best hybrid supervised-unsupervised model is chosen.

Deployment:

After comparing the output results of the developed model with the existing traditional electronic component selection method, the model is deployed on Heroku platform.

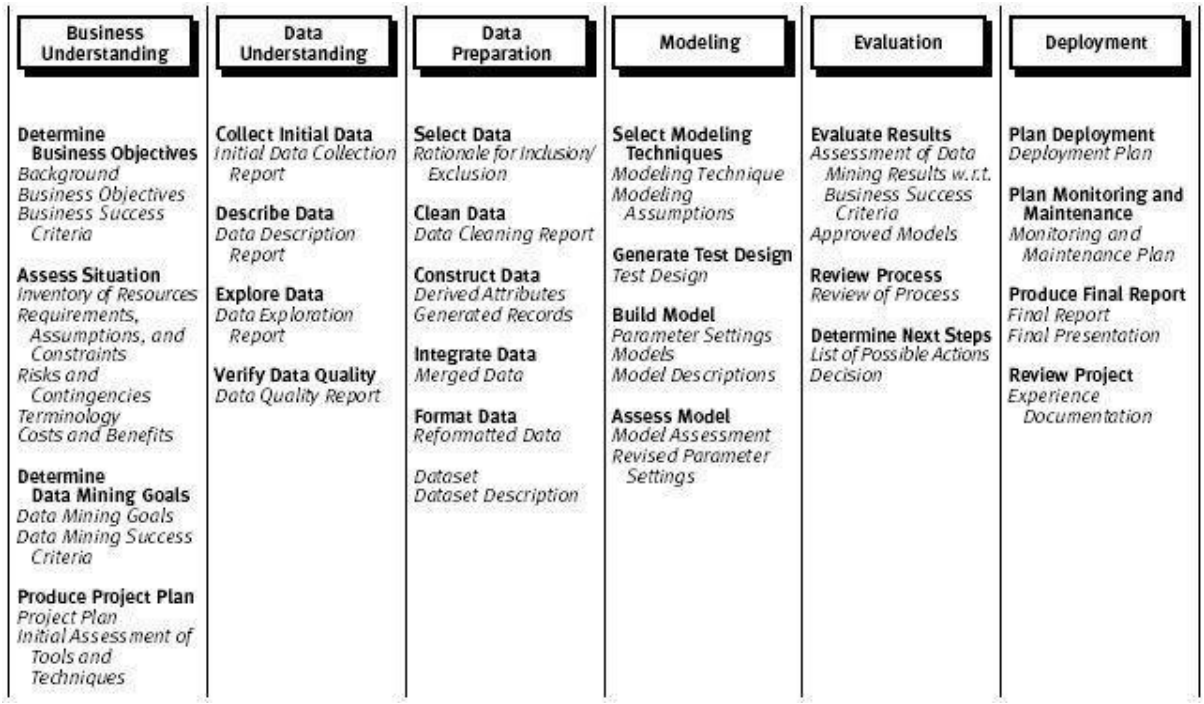

Figure 4. CRISP DM Model Phase activities [46]

3.2 Modeling Techniques

In this study, the modeling approaches to predict the best electronic component is proposed. This chapter outlines the general framework for each of the approaches. The framework is illustrated in figure 5.

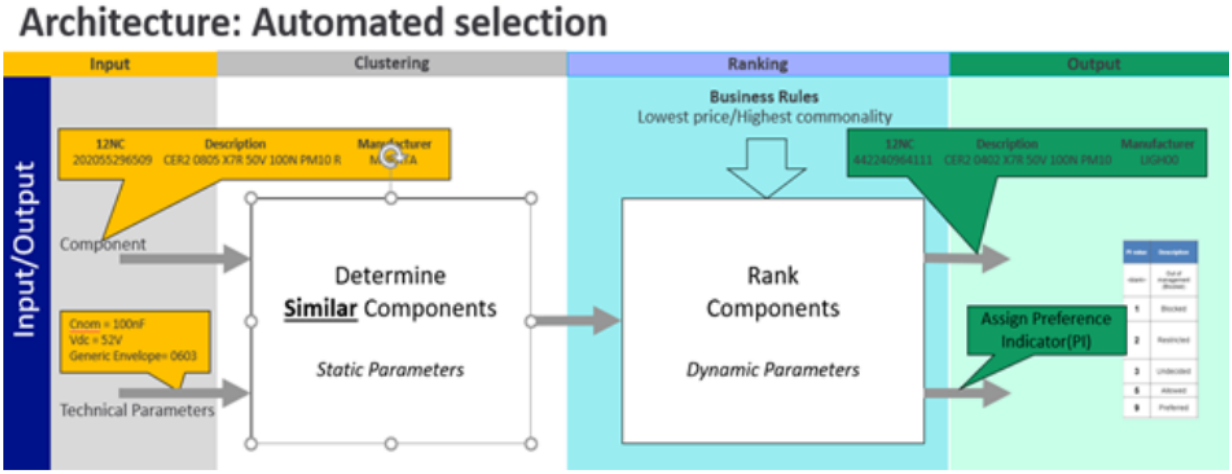

Figure 5. Proposed Model Architecture

Clustering is the technique of categorizing a bunch of things into set of interrelated items. Eventually, data items inside a cluster group are expected to exhibit a high degree of affinity to other cluster items while being significantly dissimilar from items in other clusters [47]. To arrange, sort and classify the data into different groups or cluster based on the similarity of the datapoints, various cluster algorithms are used [48]. The clustering algorithms are also used in the field of detection of outlier datapoints and recommender system [49]. Different clustering techniques are partitioned clustering, hierarchical clustering, density-based clustering etc. According to [49], the usability of various clustering techniques has been discovered to be reliant on the input data. After performing the clustering technique and obtaining the desired cluster groups based on technical properties of the component, the datapoints are ranked in order of mentioned business rules by the company. The business rule are lowest price, lowest lead time and highest component commonality. Based on the assessment of cluster results, the obtained cluster labels are used to train the model using different supervised learning techniques [50]. This helps model to classify the input query data and identify which cluster group has the most optimal component to be predicted for the same query.

3.3 Unsupervised Learning (Clustering)

In this study, we aim to predict the best electronic component to be used in electrical circuit design using hybrid machine learning approach. Two stage algorithm method of unsupervised-supervised learning are used for this purpose. The first aim is to cluster the similar electronic components into groups based on their technical feature using clustering. This includes technical parameters such as tolerance, voltage, capacitance value, resistance value, lead time, power. In this way, electronic components with identical properties are clustered. According to [47] vital important insights are retrieved from raw dataset using data mining technique.

Clustering algorithms are used to group similar components/datapoints into a cluster. The datapoints within the cluster should be identical and need to be intact with other datapoints within the cluster [47]. Clustering results are evaluated using the intercluster and intracluster datapoints distance. Each clustering algorithm has its own

limitation and advantages. Hence, it depends upon the data quality, input data structure and the business goal to determine which clustering algorithm to select.

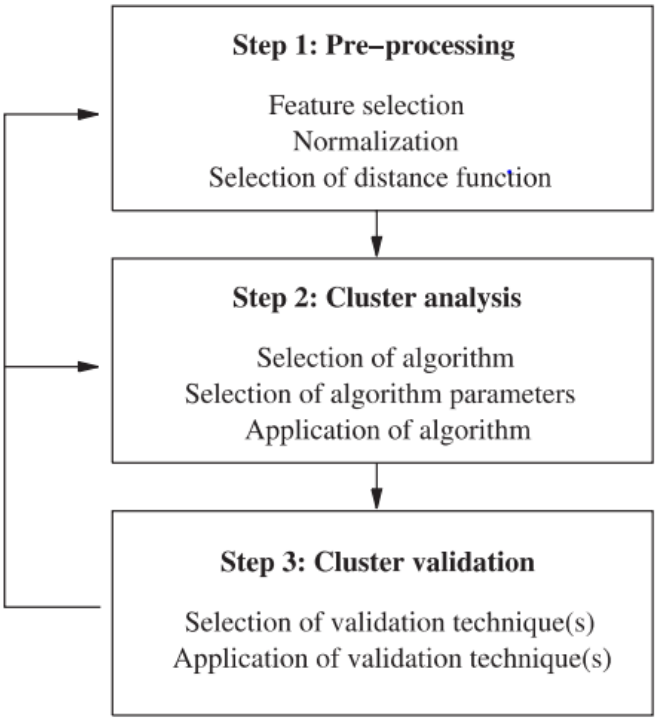

Figure 6. Steps involved in a cluster analysis [51]

3.3.1 Clustering Methods

Clustering methods such as K-means algorithm, DBSCAN, Hierarchical, BIRCH and OPTICS are few of clustering algorithms mentioned below:

K-means algorithm is an iterative centroid-based clustering algorithm. The algorithm divides the dataset into k number of distinct clusters, wherein the value of number of clusters i.e., k is predefined. In the start, each data point belongs to a nearest cluster centroid and later on based on the smallest distance between cluster centroid and datapoint, new data centers within the cluster are made. The above process is iteratively and stops as the cluster centroids are fixed. The Elbow method and silhouette score is used to measure the clusters formed. The application of this algorithm is in the field of recommendation, churn prediction, prediction and classification. This method of clustering using Expectation-Maximization approach wherein the two steps involved are assigning objects to the nearest group and

calculating the centroid of each group. The approach is described in the following steps:

1. Predefine number of clusters k.
2. Partition the dataset into predefined number of cluster k and initialize the centroid for each cluster in such a manner that one datapoint belongs to single cluster.
3. The distance between centroid and datapoint is calculated and the datapoint with smallest distance i.e., nearest datapoint is incorporated in the cluster.
4. The above process keeps on iterating until all the datapoints are distributed in the clusters and the cluster centroid stops moving.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm proposed by [52]. This method states that the density of the datapoints in a cluster plays an important role. The algorithm illustrates that the distance between the datapoints within a cluster should be the smallest and the datapoints must be quite close to each other in a cluster group. This algorithm considers two vital features viz. epsilon (minimum distance for two observations to be called as neighbor) and minPoints (minimum number of points needed to form a cluster group). DBSCAN can be used to discover clusters of any form or pattern. Firstly, all the observations in a set of data are not allocated to any cluster. The algorithm starts by selecting a random observation from the set of data that has not yet been examined. The observations are categorized as a core point if the number of observation point including this observation itself within the radius of epsilon value forms a cluster. Observation is classified as border point if it is within a accessible distance from core point and there does not exist observation points more than minPoints number of points within the vicinity. Lastly, a observation is termed as outlier if it is neither a core point nor a border point. DBSCAN algorithm helps to identify high-density datapoints region compared to less dense one.

Hierarchical clustering is an unsupervised distance based clustering algorithm. In this algorithm, a visual tree like structure is built to track the pattern of cluster splits and merge activity which is known as dendrogram. A distance matrix which contains the distances between each observation points of the dataset is generated using various distance calculation formula such as Euclidean distance, Manhattan distance

etc. There are two types of hierarchical clustering such as agglomerative and divisive hierarchical clustering [7]. In agglomerative form of clustering, all the datapoints are considered as individual cluster at first and in each iteration based on the least distances between the observation points, they are combined into a single cluster. This process continues till all the observation points are grouped into a single cluster.

Steps involved in the algorithm are as mentioned below:

1. Each data point is assigned an individual cluster.
2. Construct a distance matrix for all pairs of clusters calculating its pairwise distance.
3. Identify the pair with shortest distance and merge the two clusters.
4. Delete the entries for these clusters in the distance matrix.
5. Recalculate the distance matrix.
6. Continue repeating step 3 until step 5, until the matrix has been reduced to a single element.

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is used to deal with the problem involving huge amount of data. In this method, clustering operation is performed on a small dataset exhibiting the same properties and characteristics similar to large dataset [49] . This algorithm is referred as Two step clustering method and is a form of hierarchical clustering method. In the first stage, small cluster are formed from the large dataset with small distance threshold. In the next stage, hierarchical clustering operation is performed using the centroid of the formed clusters. In this way, a rough clustering is done to obtain tight small cluster and then iterative hierarchical clustering method is applied on this, making the overall process faster. BIRCH has an advantage of clustering multi-dimensional dataset. The cluster developed using this method is either spherical or convex in shape. It deals with only numerical values and hence categorical data is not used as input for this technique.

Ordering points to identify the clustering structure (OPTICS) helps to form clusters of varying density dataset [53] . This algorithm adds two new parameters to the DBSCAN concept such as ‘Core distance’ and ‘Reachability distance’. Core distance is the minimum value of radius needed to classify a given point as a core

point. It is the smallest value of epsilon which satisfies the minimum number of points minPts criteria to form a cluster. Next, reachability distance for a point p is described as the maximum value of either core distance or Euclidean distance. If the point is outside the epsilon radius, then reachability distance is equal to the Euclidean distance of that point. The clusters are formed using the reachability distance matrix.

Chapter 4

Data

The initial data collection plays a very crucial role in the CRISP-DM methodology's second phase. The research then moves on to getting a deeper understanding of the datasets and their contents. This is where you get your initial glimpse into the data. A strong link has been established in the business understanding step during the process.

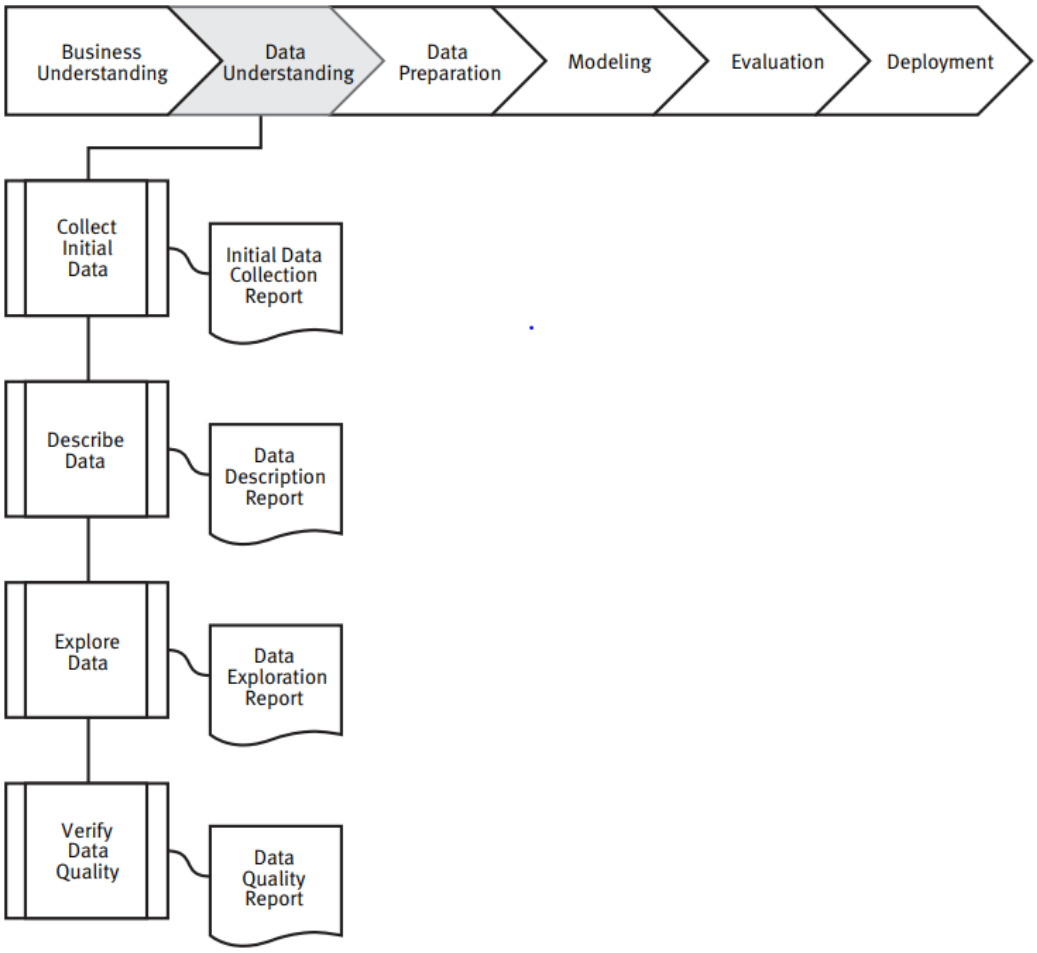

Figure 7. Data Understanding [46]

4.1 Data Gathering

Component engineering department gave a data file with electronic components information. These datasets provide information about the technical specifications and commercial parameters such as price, lead time for the component. Additionally, application and description of each component is also included in the data file which gives information about the specification of the component. Each component is assigned with a unique ‘12NC’ number which is mentioned in the dataset. The data set was extracted from Signify ’xDM‘ component database and was made available to us in ‘csv’ file format.

4.2 Data Description

The Data consisted of two datasets for Capacitor and Resistor components. The aim of this project is to design a model which predicts the best component to be used in circuit design. To fulfill this aim, we were asked to work with Capacitor and resistor dataset.

Capacitor Dataset

This dataset contains 2,30,000 entries of different types of capacitors being supplied to Signify through various suppliers. Based on the technical needs of the design, a designer/engineer selects the best considering technical aspects, but commercial business factors are missed by the engineer. Provided data file contains the below information per entry:

12NC: A unique component id number.

Description: Technical description of the component.

Manufacturer Code: Code for specific component manufacturer.

Generic Envelope: Categorical data for specific components.

C-nom: Capacitance value measured in farad.

U_Rdc: Rated Voltage.

Upper and Lower Tolerance: Capacitance tolerance range.

MOQ: Minimum Order Quantity.

Price: Price of the component.

Lead Time: Measured in No. of days.

Application: Either for General purpose or precision purpose.

Commonality_table: This unit specifies the number of circuits in which this type of component can be used. (Higher the commonality better for the company).

Resistor Dataset

This dataset contains 2,85,000 entries of different types of resistors being supplied to Signify through various suppliers. Based on the technical needs of the design, a designer/engineer selects the best considering technical aspects, but commercial business factors are missed by the engineer. Provided data file contains the below information per entry:

12NC: A unique component id number.

Description: Technical description of the component.

Manufacturer Code: Code for specific component manufacturer.

Generic Envelope: Categorical data for specific components.

R-nom: Resistance value measured in ohm.

P_max: Resistor Power rating, it's the maximum quantity of heat released indefinitely by a resistive component without diminishing its efficiency.

Upper and Lower Tolerance: Resistance tolerance range.

MOQ: Minimum Order Quantity.

Component Price: Price of the component.

Lead Time: Measured in Number of days.

Application: Either for General purpose or precision purpose.

Commonality: This unit specifies the number of circuits in which this type of component can be used. (Higher the commonality better for the company).

4.3 Data Preparation for Cluster Algorithms

In the clustering process, [54] emphasize the need of normalization and standardization. Because of the varied units of measurement, the values of each property can vary in ranges. This raises the possibility that some characteristics will outweigh others. The outcomes of the clustering procedure are greatly influenced by data normalization and standardization. The clustering procedure's findings are highly impacted by data. However, research has revealed that Normalization isn't necessarily a good thing. The source of data recording must be determined by researchers and assess if it is necessary to normalize data or if doing so makes the data weaker.

[55] recommend that the parameters used in the cluster analysis be chosen with caution. It is not advised for researchers to employ several factors accessible. The right parameters for the research should be picked, as choosing any unrelated parameter could impact clustering results.

[56] points out that normalization can be done in a variety of ways. On the original data, the Min-Max Normalization applies a linear change.

4.3.1 Data Selection

From the capacitor data file, the features '12NC', 'commonality_table', 'Generic Envelope', 'C-nom', 'U_Rdc' and 'Price' are considered to form a DataFrame and the rest are discarded. The Generic Envelope attribute is filtered for only four categories electronic components viz. ‘0402’, ‘0603’, ‘0805’,’1206’ as the company deals with these categories lately. Commonality_table consists of a value which refers to the number of circuit designs in which the selected component can be used for. This helps companies to select the component with highest commonality first as they can buy the component from the supplier in large quantities at a discount rate. The missing values

in ‘commonality_table’ attribute is filled with zero value assuming they have never been selected for any design. The missing values in the ‘component price’ attribute is filled using mean imputation methods. After imputing missing values in the price column, rows with missing data for '12NC', 'Generic Envelope', 'C-nom', 'U_Rdc' are dropped as these are technical specification data of the component and cannot be imputed.

From the resistor data file, the features '12NC', 'commonality_table', 'Generic Envelope', 'R-nom', 'P_max ', 'Component Price' are considered to form a DataFrame and the rest are discarded. Commonality_table consists of a value which refers to the number of circuit designs in which the selected component can be used for. This helps companies to select the component with highest commonality first as they can buy the component from the supplier in large quantities at a discount rate. The missing values in ‘commonality_table’ attribute is filled with zero value assuming they have never been selected for any design. The missing values in the ‘component price’ attribute are filled using mean imputation methods. After imputing missing values in the price column, rows with missing data for '12NC', 'Generic Envelope', 'R-nom', 'P_max are dropped as these are technical specification data of the component and cannot be imputed.

4.3.2 Missing Value Imputation

In this study, eliminating rows having any of the feature value missing would be a poor solution because it would drastically limit the sample size, perhaps resulting in high standard errors. As a result, in this circumstance, the use of imputation methods is recommended. There are a variety of approaches that can be used to address this problem [57]. The following basic, yet effective strategy given by [58] is for calculating and evaluating the effectiveness of various methods. The Imputation methods used in this study are Mean Imputation and k-Nearest Neighbors Imputation.

1. In the mean imputation method, data for the missing feature value is imputed by calculating mean value for that feature values in the dataset.
2. We came to the conclusion that none of the entries in the two datasets have an odd or improbable value after evaluating all of them and also no duplicate values were found in the dataset.

4.3.3 Data Normalization

Though most of data characteristics belong to the same type of data, their range of values differ. After all the data features have been chosen, the dataset is ready to be clustered. Because clustering methods construct group of clusters depending on feature distances, it is important to bring all the features to a standard scale value. Table 4. illustrates the value range for each data features.

Table 4. Value ranges of Data Features

Data Features	Value range
R-nom	[0 ; 1.500000e+07]
R_Tol	[0 ; 20]
P_max	[0.05 ; 0]
Commonality	[0 ; 1246]
Price	[0.0011 ; 0.91]
Lead Time	[0 ; 40]
C-nom	[1.0e-13 ; 4.7e-5]
U_Rdc	[6.3 ; 6000]

If this is kept untouched, clustering the data will be difficult. Because certain value ranges are much bigger than others, perhaps by a factor of 150.000 or more. The attribute with the larger value range could dominate the distance calculations between the data points computed the clustering algorithms, to avoid this scenario normalization techniques are used. Normalization can be accomplished in a variety of methods. The min-max normalization technique is utilized in the study. This method's mathematical formula is:

$$Normalized\ value = \frac{\text{selected value} - \text{minimum value}}{\text{maximum value} - \text{minimum value}}$$

4.4 Conclusion

The data features which are determined to be significant in both the dataset of capacitor and resistor are transformed into a new DataFrame and fed to the clustering models. C-nom, U_Rdc, Generic Envelope, Price and commonality are the attributes

chosen for capacitor dataset. The data attributes chosen for the resistor dataset are R-nom, P_max, Generic Envelope, Price and commonality.

Min-max normalization data transformation method is used to scale values in the range of 0 to 1, to minimize the error in computation of distances between data points while using clustering algorithms.

Data Entries with missing values for price and C-nom attribute were handled using mean Imputation and K-nearest neighbor imputation techniques. The entries not having a unique ‘12NC’ component id number were dropped as imputation methods cannot be used for it.

Chapter 5

Model Development

In this chapter, hybrid model using unsupervised and supervised machine learning technique is built. The model design assumptions and parameter settings are performed in this stage.

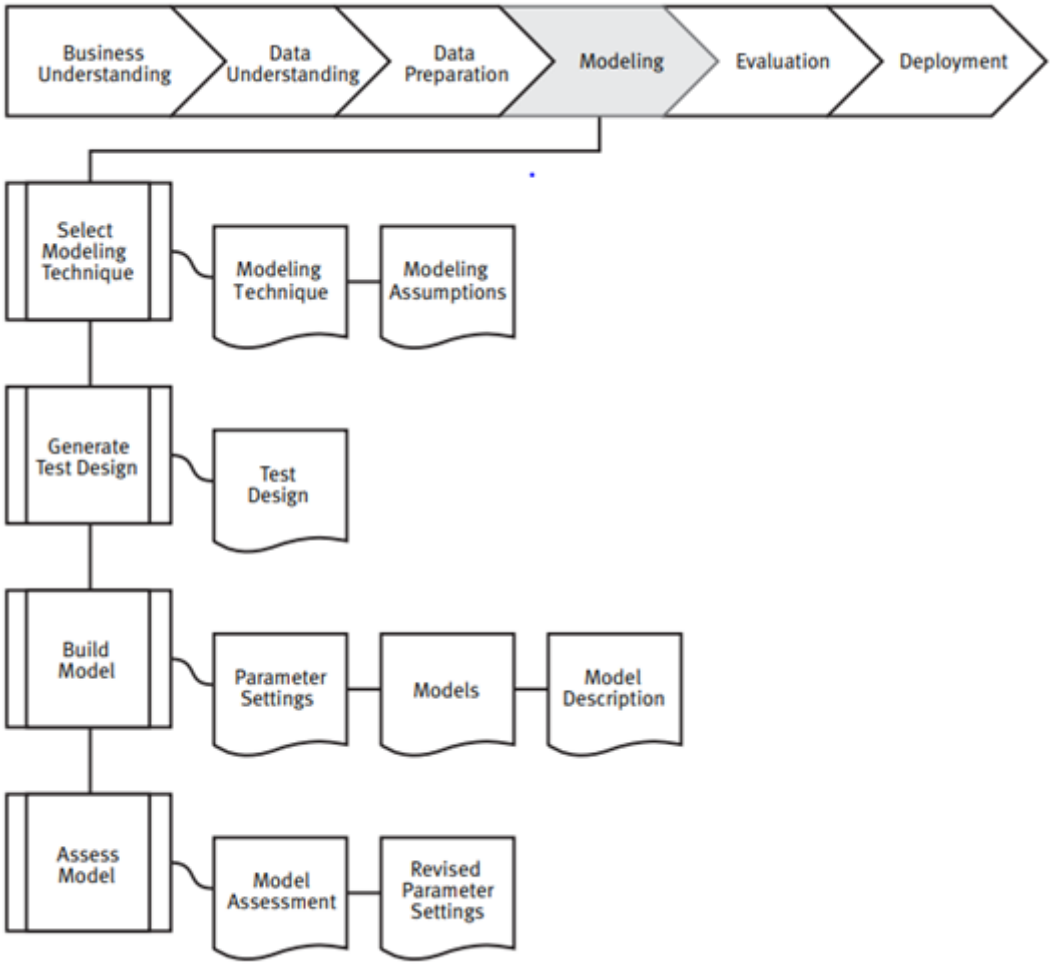

Figure 8. CRISP-DM Modelling Phase [46]

The data prepared in the previous section must be fed into the different clustering algorithm to assess the result.

5.1 Model Selection

Unsupervised Learning (Clustering)

Based on the literature review findings, five clustering algorithms have been chosen for implementation out of all the cluster approaches examined. These are Hierarchical Agglomerative clustering, DBSCAN, BIRCH, Mean Shift and OPTICS.

Hierarchical clustering is an unsupervised distance based clustering algorithm. In this algorithm, a visual tree like structure is built to track the pattern of cluster splits and merge activity which is known as dendrogram. A distance matrix which contains the distances between each observation points of the dataset is generated using various distance calculation formula such as Euclidean distance, Manhattan distance etc. There are two types of hierarchical clustering such as agglomerative and divisive hierarchical clustering [7]. In agglomerative form of clustering, all the datapoints are considered as individual cluster at first and in each iteration based on the least distances between the observation points, they are combined into a single cluster. This process continues till all the observation points are grouped into a single cluster.

DBSCAN method states that the density of the datapoints in a cluster plays an important role. The number of clusters as input is not needed for this method. The algorithm illustrates that the distance between the datapoints within a cluster should be the smallest and the datapoints must be quite close to each other in a cluster group. This algorithm considers two vital features viz. epsilon (minimum distance for two observations to be called as neighbor) and minPoints (minimum number of points needed to form a cluster group). DBSCAN can be used to discover clusters of any form or pattern. Firstly, all the observations in a set of data are not allocated to any cluster. The algorithm starts by selecting a random observation from the set of data that has not yet been examined. The observations are categorized as a core point if the number of observation point including this observation itself within the radius of epsilon value forms a cluster. Observation is classified as border point if it is within a accessible distance from core point and there does not exist observation points more than minPoints number of points within the vicinity. Lastly, an observation is termed as outlier if it is neither a core point nor a border point. DBSCAN algorithm helps to

identify high-density datapoints region compared to less dense one [17]. This technique is also robust to outliers and arbitrary shape clusters are built using it.

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is used to deal with the problem involving huge amount of data. In this method, clustering operation is performed on a small dataset exhibiting the same properties and characteristics similar to large dataset [49] . This algorithm is referred as Two step clustering method and is a form of hierarchical clustering method. In the first stage, small cluster are formed from the large dataset with small distance threshold. In the next stage, hierarchical clustering operation is performed using the centroid of the formed clusters. In this way, a rough clustering is done to obtain tight small cluster and then iterative hierarchical clustering method is applied on this, making the overall process faster. BIRCH has an advantage of clustering multi-dimensional dataset. The cluster developed using this method is either spherical or convex in shape. It deals with only numerical values and hence categorical data is not used as input for this technique.

Ordering points to identify the clustering structure (OPTICS) helps to form clusters of varying density dataset [53] . This algorithm adds two new parameters to the DBSCAN concept such as ‘Core distance’ and ‘Reachability distance’. Core distance is the minimum value of radius needed to classify a given point as a core point. It is the smallest value of epsilon which satisfies the minimum number of points minPts criteria to form a cluster. Next, reachability distance for a point p is described as the maximum value of either core distance or Euclidean distance. If the point is outside the epsilon radius, then reachability distance is equal to the Euclidean distance of that point. The clusters are formed using the reachability distance matrix.

Mean Shift Clustering is a type of hierarchical clustering method. In this approach, every feature set is considered as a cluster center. Unlike other partitional clustering algorithm, in this method there is no need to specify the number of clusters prior to training. A bandwidth is formed across each datapoints and within the bandwidth all the feature set is considered. The mean distance of all the datapoints within the bandwidth is calculated and the new cluster center with new bandwidth is formed. In the next step, for the cluster again all the features are selected, and the mean distance is formed. This process keeps on iterating until the bandwidth center is fixed and no

new centers are formed. At this stage dataset is clustered into desired number of clusters using this algorithm.

Clustering algorithm helps to group together similar and homogeneous components in capacitor and resistor dataset in clusters considering technical parameters such as capacitance value C-nom, Voltage (U_Rdc) for capacitor and R-nom, P_max for resistor dataset. The performance of these algorithms is measured based on metrics such as Silhouette score, Calinski-Harabasz (CH) score, David-Bouldin (DB) score.

In our approach, we are not using K-means clustering as we are not able to estimate the initial number of clusters, which is a significant factor for K-means. Hence, we have used a distance threshold parameter for the above 5 mentioned algorithms. In this, we have calculated a distance matrix which computes the distance between all data objects. Once we have computed the Euclidean distance between each data point (components in the dataset), based on the distance threshold value, the cluster is formed. For example, if the distance threshold value is set to 0.15, then all components having value less than or equal to 0.15 will be in one cluster. In this way, without providing an initial number of clusters we are able to get similar components clustered together. The components are ranked in a cluster as per business rule which is low price, high commonality and load lead time.

Once the clusters are formed, a supervised learning model is trained on the obtained clusters as labels. In this way, the model will be able to classify a query from the user based on technical inputs to the suitable formed cluster to recommend the best component for that input query.

Supervised Learning

A comprehensive list of Supervised learning algorithms is explored, and five models are finalized based on literature review findings. The algorithms used are K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest Classifier (RF), XGBoost Classifier and Gaussian Naive Bayes (NB).

K-Nearest Neighbors algorithm (KNN) is an algorithm based on supervised learning techniques, mainly used for classification and regression problems. KNN does not make calculations or assumptions on the underlying data, hence is a non-parametric

algorithm. It works by selecting the k nearest neighbors based on Euclidian distance and classifies the data point to a particular class based on the classes of the k neighbors considering the majority vote on neighbors classes.

Support Vector Machine classifier is a Supervised Learning algorithm, used for Classification and Regression problems. In supervised learning domain, it is very prominent algorithm. It works by dividing the multi-dimensional space into classes. It creates these boundaries to classify data points in the space. There can be multiple decision boundaries, the best of which is referred to as a hyperplane. Support Vectors are the closest data point to the optimal decision boundary (hyperplane).

Random Forest is a machine learning algorithm that is built on the principle of ensemble learning, in which multiple classifiers are employed in the resolution of a bigger complex problem. In supervised learning domain, it is a very prominent algorithm. It works by creating multiple decision trees based on random data points selected and uses the average to increase the dataset's prediction accuracy. For prediction the algorithm takes input in the form of predicted results from multiple decision trees and does its final prediction based on the majority of the votes.

XGBoost, is a popular supervised machine learning algorithm, famous for its exceptional computational speed. It is built on gradient boosting decision trees algorithm. It works by creating new models sequentially. In each sequence the error from the previous model is reduced until there is no more room for improvement, in this way it minimizes the total loss for prediction/classification problems.

Naive Bayes is a supervised leaning algorithm mostly used for text classification. It is built on Bayes theorem, on the concept of conditional probability. Naive Bayes algorithm performs exceptional for high dimensional training dataset, like text classification and can be used for both binary and multi class classification.

5.2 Model Result Validation

The primary goal of evaluation indicators is to determine whether or not an algorithm is valid Clustering techniques are built on the premise of distance, also known as dissimilarity, and similarity [59]. When dealing with computable data characteristics,

detecting correlations between data attributes is calculated using distances, however when dealing with qualitative data attributes, similarity is desired.

Clustering, according to [60] is exploratory rather than confirmatory. The clustering results appear to be natural if each cluster is dense and separated from the others. Cluster validity, they claim, provides some assurance that the cluster structure discovered is meaningful.

As [61] concluded, validation measures deliver large volumes of information which would not be possible to gather solely by eye inspection. There are a variety of validation methods available and using a combination of them helps reduce the chance of misinterpreting data. This method, it is determined, maximizes the confidence in the produced results. [62] agrees that employing various validity indices is a good idea.

Several cluster validation indices have been discovered. These can be classified as either internal or external. **Davies-Bouldin (DB) index, CalinskiHarabasz (CH) score and silhouette score** are the most commonly used validity indices, according to research [63]. All this validation index uses Euclidean distance between the data as an objective function.

For Supervised classification learning, models are using Accuracy, Precision_score, Recall, Cohen_Kappa score and f1-score measure score to access the effectiveness of classification models, as stated by [60].

Accuracy is a prominent metric in classification algorithms especially for multi class classification. Is it derived from the confusion matrix. It is calculated by the ration of the sum of all True Positive (TP) and True Negative (TN) to all entries in confusion matrix added together. In a confusion matrix, TP and TN are the data points that were correctly classified, and they lie on the diagonal.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The Precision, also known as positive predictive value, is the ratio of the number of True Positives (TP) to the sum of True Positives (TP) and False Positives (FP). TP

occurs when the model predicts the positive class correctly, FP occurs when the model predicts Positive class incorrectly. The Precision determines the proportion of positives predictions that were actually correct.

$$Precision = \frac{TP}{TP + FP}$$

The Recall, also known as true positive rate, is the ratio of the number of True Positives (TP) to the sum of True Positives (TP) and False Negatives (FN). TP occurs when the model predicts the positive class correctly, FN occurs when the model predicts negative class incorrectly. The Recall determines the proportion of actual positives that were predicted correctly.

$$Recall = \frac{TP}{TP + FN}$$

The **Cohen kappa score** determines the level of reliability of two raters and its mutual exclusiveness. It measures how after the raters agree.

F1 Score is metric which is computed by taking harmonic mean of the Precision (P) and the Recall (R) of the classification model. Both false positives and false negatives are considered in this score.

F1 Score = 2 * (P * R) / (P + R)

Chapter 6

Result Analysis

The primary findings from the modeling process are presented in this chapter. Firstly, results for unsupervised learning approach i.e., cluster modeling analysis process are presented in Section 6.1. Then, the results of trained classification models are evaluated according to the metrics mentioned in the previous section. After evaluating the results for both supervised-unsupervised models, the best performing one amongst the two are selected to develop the hybrid machine learning model for predicting the best component.

6.1 Result Analysis of Cluster Modeling approach

Cluster validation will be divided into two parts. Validation measures are used to assess the cluster algorithm's ability to separate groups based on density and distance. Expert opinions and knowledge from component engineers are utilized to validate the outcomes. Several cluster validation indices have been discovered. These can be classified as either internal or external. Three internal validity indicators are used in this project: the Silhouette score, the DB index and the CH score. The three metrics' validation results are displayed below in Table 5. for resistor dataset.

For Resistor Dataset:

Table 5. Cluster validation metrics for resistor

Validity Index	Hierarchical Agglomerative	DBSCAN	BIRCH	Mean_ Shift	OPTICS
Silhouette score.	**0.9237**	0.094	0.9145	0.913	0.7496
Davies-Bouldin	0.69	0.3224	0.4240	0.08	**0.7256**
Calinski-Harabasz	61.72	6.95	**295.88**	1.75	137.65

From the above table we observe that the DB index value is lowest for Mean_Shift compared to the other four clustering algorithms. DB index value represents the ratio of the intra cluster distances between data points to the inter cluster distances. As the

value of DB index for Mean_Shift clustering algorithm is the lowest, it signifies that the clustering results are better compared to other methods. CH index value represents the similarity of a data point to its own cluster in comparison to the other clusters. For BIRCH the CH value is high, which signifies the cluster are dense and well distributed. Silhouette Score is a function that defines the ratio of difference between average of intra cluster distances to its average inter cluster distance for each data point. Hierarchical agglomerative clustering outperforms other clustering methods with respect to Silhouette score. In certain ways, all three approaches are better than the others, making it challenging to choose the optimal algorithm. The result of clustered components obtained by applying all 5 algorithms were exported in excel file format and compared with the technical expertise knowledge team of signify to confirm which clustering algorithm is giving the desired result. A deviation percent of 10% for resistance (R-nom) and power (P_max) i.e., components within 10% of its range of value were clustered together. For e.g., a component with specification of resistance (R-nom) value as 100 ohm and power (P_max) value of 50 W will be clustered with component having R-nom not greater than 110 ohm and not less than 90 ohm, also P_max not greater than 55 W or less than 45 W. This was the requirement mentioned by the company for cluster validation.

Based on the analysis of results using exported files, Hierarchical agglomerative clustering was yielding the best cluster result with the highest silhouette score amongst all other algorithms. After the components are clustered into clusters based on features such as Resistance value (R-nom) and Power (P_max), the next step was to rank the components in each cluster as per specified business rule of the company. The rule is to first recommend or select the component with the lowest price within the cluster as per user query. If the prices are similar within the cluster for the component, then the component with lowest Generic Envelope value needs to be recommended. At the end, if the price and Generic Envelope are similar for components within a cluster then the component with highest commonality value should be recommended to the user. These are the business rules enlisted by the company which are needed to be followed for ranking the component in the same cluster for both datasets. On successful ranking of the component within each cluster based on commercial business rule, supervised learning approach is used to train the model with

cluster_labels as classes. In this way, it makes our classification approach a multi-class classification method.

Analysis of Result for Supervised learning Classification approach:

The analysis of five supervised ML classification techniques are evaluated using metrics such as Accuracy, Precision, Cohen_Kappa score, Recall score and F1-score.

Table 6. Result for supervised learning resistor dataset

Metric	Random Forest	KNN	SVM	Naive Bayes	XGBoost
Accuracy	0.9480	**0.9531**	0.3188	0.1704	0.9282
Precision_Macro	0.874	0.8556	0.1626	0.06	0.7792
Precision_Micro	0.9480	0.9531	0.3188	0.1704	0.9282
Precision_weighted	0.9459	**0.9505**	0.2093	0.103	0.9081
Cohen_Kappa	0.9479	**0.9529**	0.3168	0.168	0.9280
Recall	0.9480	**0.9531**	0.3188	0.1704	0.9282
F1-score	0.9457	**0.9486**	0.222	0.108	0.9131

Amongst all the classification model used, **K-Nearest Neighbor classifier** performs better with an accuracy score of 0.9531, precision score of 0.9505, Cohen_kappa score of 0.9529, Recall score of 0.9531 and f1 score of 0.9486.

For Capacitor Dataset:

The same approach as mentioned for Resistor dataset is followed for capacitor dataset with only features such as R-nom replaced by C-nom (Capacitance value) and P_max replaced by U_Rdc (voltage) for clustering purpose. The business rule used for ranking of components after clustering is the same for this dataset also. The three metrics' validation results are displayed below in Table 7. For capacitor dataset.

Table 7. Cluster validation metric for capacitor

Validity Index	Hierarchical Agglomerative	DBSCAN	BIRCH	Mean Shift	OPTICS
Silhouette score.	**0.9342**	0.9166	0.8143	0.8	0.8245
Davies-Bouldin	0.6982	**0.07**	0.6625	0.5402	0.4355
Calinski-Harabasz	**2132.75**	50.29	360.49	315.48	2046.19

From the above table we observe that the DB index value is lowest for DBSCAN compared to the other four clustering algorithms. DB index value represents the ratio

of the intra cluster distances between data points to the inter cluster distances. As the value of DB index for DBSCAN clustering algorithm is the lowest, it signifies that the clustering results are better compared to other methods. CH index value represents the similarity of a data point to its own cluster in comparison to the other clusters. For Hierarchical Agglomerative clustering the CH value is high, which signifies the cluster are dense and well distributed. Silhouette Score is a function that defines the ratio of difference between average of intra cluster distances to its average inter cluster distance for each data point. Hierarchical agglomerative clustering outperforms other clustering methods with respect to Silhouette score. In certain ways, all three approaches are better than the others, making it challenging to choose the optimal algorithm. The result of clustered components obtained by applying all 5 algorithms were exported in excel file format and compared with the technical expertise knowledge team of signify to confirm which clustering algorithm is giving the desired result. A deviation percent of 10% for capacitance (C-nom) and voltage (U_Rdc) i.e. components within 10% of its range of value were clustered together. For e.g., a component with specification of Capacitance (C-nom) value as 100nf and Voltage (U_Rdc) value of 50V will be clustered with component having C-nom not greater than 110 nf and not less than 90nf, also U_Rdc not greater than 55V or less than 45V. Based on the analysis of clustered results using exported files, **Hierarchical agglomerative clustering was yielding the best cluster result with the highest silhouette score amongst all other algorithms.** After the components are clustered into clusters based on features such as capacitance value (C-nom) and U_Rdc (Voltage), the next step was to rank the components in each cluster as per specified business rule of the company mentioned earlier. On successful ranking of the component within each cluster based on commercial business rule, supervised learning approach is used to train the model with cluster_labels as classes. In this way, it makes our classification approach a multi-class classification method.

The analysis of five supervised ML classification techniques are evaluated using metrics such as Accuracy, Precision, Cohen_Kappa score, Recall score and F1-score. The results are shown in below table. Amongst all the classification model used, **K-Nearest Neighbor classifier** performs better with an accuracy score of 0.9221, precision score of 0.9074, Cohen_kappa score of 0.9213, Recall score of 0.9221 and f1 score of 0.9128.

Table 8. Result of supervised learning algorithm for capacitor

Metric	Random Forest	KNN	SVM	Naive Bayes	XGBoost
Accuracy	0.2305	**0.9221**	0.0568	0.0658	0.8383
Precision_Macro	0.1526	0.7636	0.0046	0.0060	0.6098
Precision_Micro	0.2305	0.9221	0.0568	0.0658	0.8383
Precision_weighted	0.1866	**0.9074**	0.0058	0.0089	0.7663
Cohen_Kappa	0.2232	**0.9213**	0.0476	0.5617	0.8365
Recall	0.2305	**0.9221**	0.0568	0.0658	0.8383
F1-score	0.1921	**0.9128**	0.0103	0.0152	0.7935

Considering the results of various supervised-unsupervised models, we came up with a combination of a hybrid unsupervised-supervised learning approach to predict the best electronic component. Firstly, using an unsupervised learning technique of Hierarchical agglomerative clustering, the similar components are grouped together based on technical parameters. After the clustering is done, ranking of components based on business rules is performed on each cluster. In this way the components are sorted with the lowest price component being ranked at the top. This approach is needed as in the current traditional method also company follow the same business rule for comparing the component and recommending the best one. After ranking and obtaining the optimal number of clusters, the model is trained using supervised learning approach to classify new input queries. For this K-nearest neighbor is used as it outperformed other models with higher accuracy score and precision score. **In this way, a combined hybrid model of supervised-unsupervised learning is used to predict the best electronic component for design purposes.**

6.2 Model Result Validation

In this section, the outcomes of the generated ML model are compared with the existing traditional method of component recommendation to gain insights on the effectiveness of the proposed hybrid ML model.

For validation purposes, a test dataset was provided by the Company which consisted of technical parameters such as ‘12NC’, ’C-nom’, ’U_Rdc’, ‘R-nom', ’P_max’, ‘Component description’ and we were asked to predict the best component for the entries based on their technical features (’C-nom’, ’U_Rdc’ for capacitor and ‘R-nom’, ’P_max’ for resistor). The best components for the entries in the shared test dataset

were already predicted by the company a month ago using their traditional manual method. Hence, the results were to be validated with our model predictions to measure prediction accuracy of proposed model. The test dataset is utilized as input to the developed model and results in the form of 'predicted components' were added as a new column in the dataset. The model output of predicted components was compared with the existing predicted component result done by the company using a manual process to measure prediction accuracy. An accuracy of 84% for capacitor and 81% for resistor was achieved using the proposed ML model which means 84% of the number of capacitor components were predicted correctly by the model considering their technical and commercial parameters.

6.3 Model Deployment

Based on the result obtained from application of hybrid supervised - unsupervised machine learning methods for both the dataset, Hierarchical agglomerative clustering and K-nearest neighbor algorithms are used to develop a hybrid ML model for prediction of best electronic component. Using the python pickle library, the model is serialized and dumped into a **.pkl file.**

A lightweight web application is built using Python web framework 'Flask'. Flask is web framework used in python to create web applications. We have designed a web application in which the design engineer can query based on input technical attribute values such as c-nom and U_Rdc for capacitor dataset and R-nom and P_max for resistor dataset which are used as input to the model. The model will predict/recommend the optimal electronic component to the design engineer based on the training provided to it. When a user queries through the page, the web application page displays the top 5 best ranked predicted components with their technical details in a tabular form.

The model is deployed on Heroku platform which is an infra as a service platform used to deploy and scale software and web engineering applications [64].

Below image displays the webpage wherein a user needs to input two of the technical parameters to get the top 5 best components.

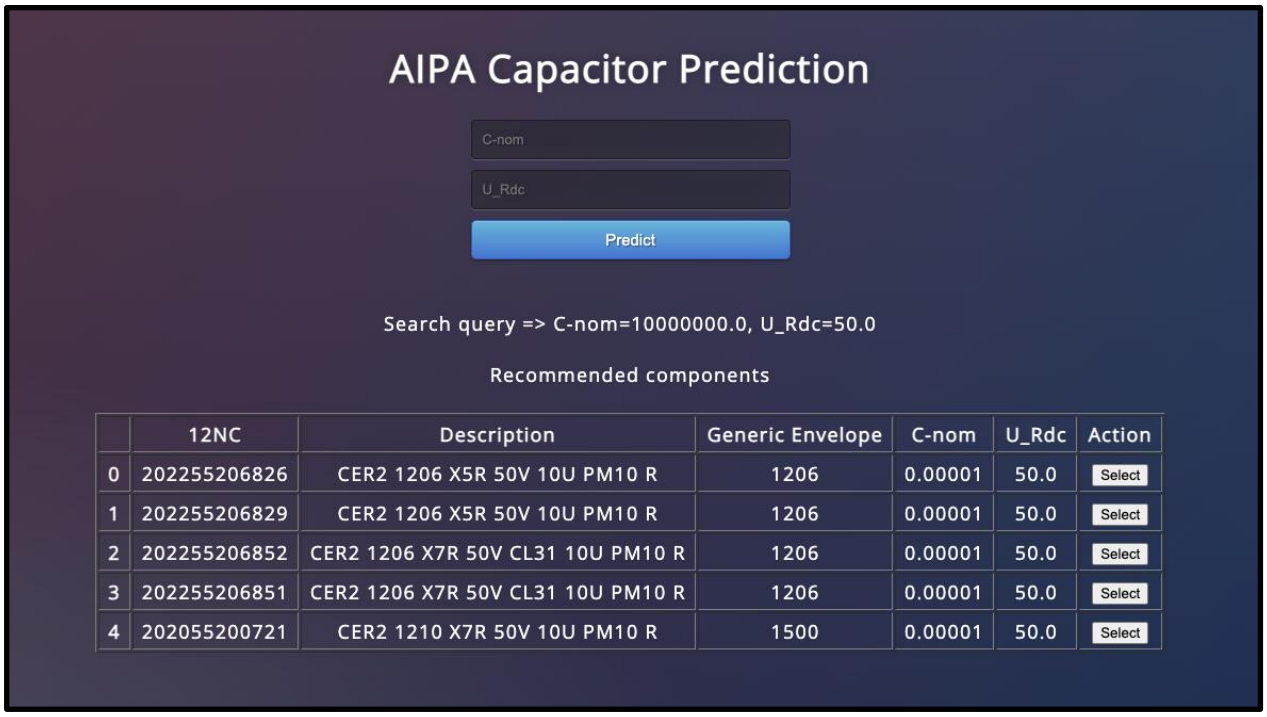

Figure 9. Capacitor model webpage

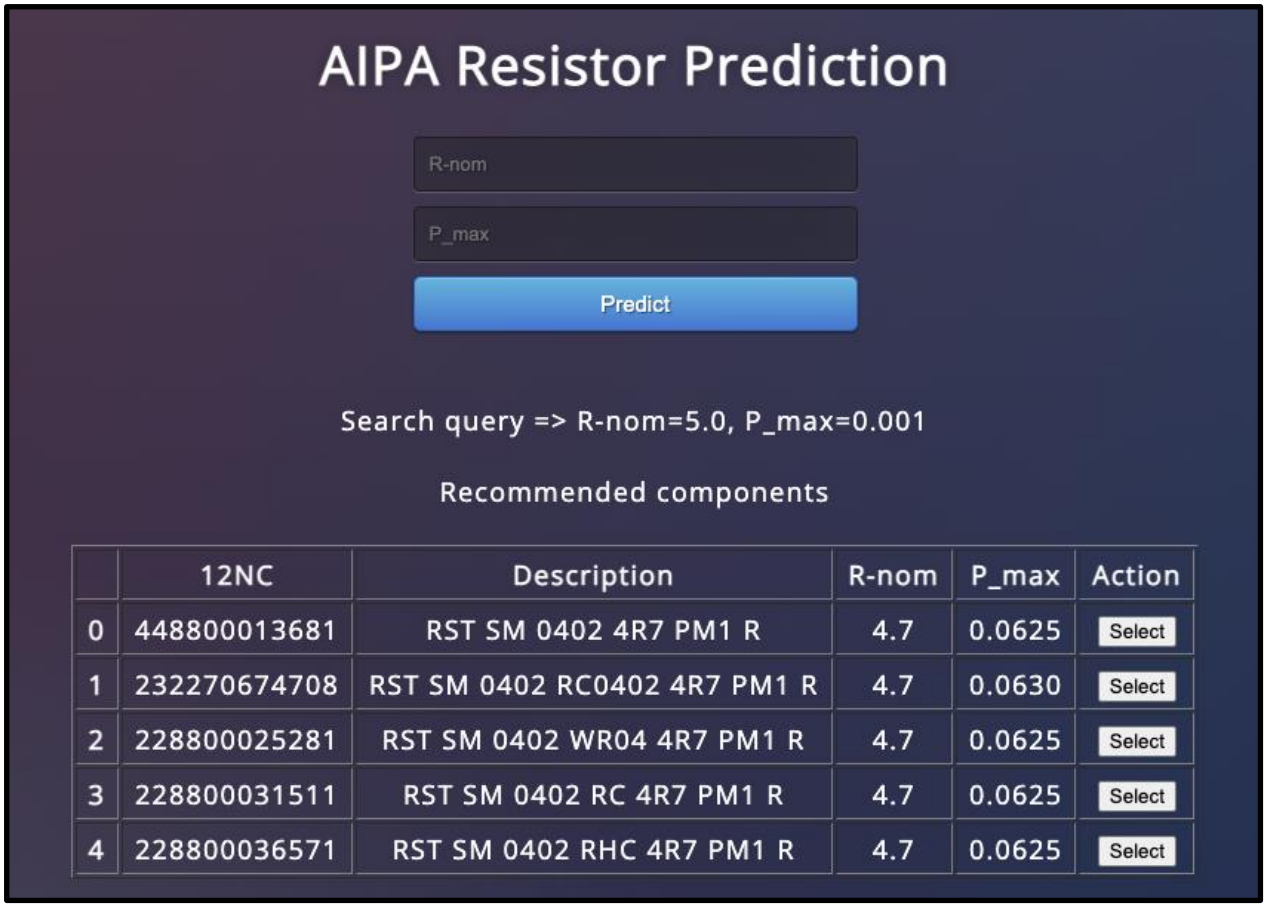

Figure 10. Resistor model webpage

The webapp link for capacitor and resistor is mentioned below:

1. Capacitor - https://aipa-signify.herokuapp.com/

2. Resistor - https://aipa-signify.herokuapp.com/resistor

Chapter 7

Conclusion

The aim of the study is to evaluate the feasibility of using machine learning techniques to automate the electronic component selection process at Signify. Automating the process dissects with the grouping of similar technical characteristic components together to form a cluster using an unsupervised machine learning approach and then applying supervised machine learning methods to train the model using cluster labels as classes. Related to these goals the main research question is formed which states to ‘Evaluate the value of combining supervised-unsupervised machine learning to predict the best electronic component’. The motivation behind the proposed hybrid supervised-unsupervised modeling technique needs to be discussed to gain more insights of the effectiveness of the proposed approach.

7.1 Motivations of Automated electronic component selection

A company that wants to be successful must be efficient in its product delivery process. The compatibility of cross-functional teams is a significant aspect in the company's success. The organization benefits from good collaboration across engineering design, procurement, production, and business development teams. Component selection is the process of choosing an appropriate component or a set of similar components from a variety of supplier to enable the designed electrical circuit to perform as intended [3]. Because they serve as a link between the electrical design engineer, material procurement, and product assembly/manufacturing units, the component engineering team is responsible for optimal component selection. The supplier's selection of the proper component has an impact on a device's performance and consistency. A well-designed technique for predicting the most ideal electrical component to use in design might help to shortlist number of potential suppliers. This process needs to be scalable to meet the component demand variability requirement.

Hence, the performance of a component engineer is very vital for the progress of a company's business in the semiconductor industry. This process of electronic component selection/recommendation is traditionally performed by a team of component engineering experts and the decision is made through manual human judgement and knowledge. As this process is prone to human error and also time consuming, there is a need to automate the process and evaluate its efficacy compared to existing method in the semiconductor industry. Automated process would be able to solve the problems such as unpredictable workload of component selection, providing substitutes for deprecated and high-cost price components and saving cost and-time for the company. Hence, developing a model which combines supervised and unsupervised machine learning will help to reduce the existing process complexity in the component selection process, can provide cheaper substitute component to be used in design from a huge database in a flash of seconds and would also help to reduce the overall product delivery time for the company which in turn will help in revenue generation. The process of electrical circuit designs moving to and fro for approval can be resolved through the model, predicting the best component. This also would reduce the waiting time for design to get approved through a component engineering team.

7.2 Research Question Answers

RQ1: Which of the existing machine learning approaches could be used to predict the best electronic component?

Based on the Literature review study, many of the existing component selection techniques use an aggregation approach for determining a recommendation [65]. Study by [66] presents a hybrid technique, combining simulation and machine learning and examines its applications to data-driven decision-making support in resilient supplier selection.

A study by [22] presents a hierarchical clustering-based method to solve a supplier selection problem and find the proximity of the suppliers. Another study by [41] also concluded that clustering the components based on their parameters similarities to

form clusters and then applying C4.5 Decision tree classifier algorithm performs better than traditional manual component selection.

Hybrid data mining models, where unsupervised and supervised learning techniques are combined, are proven to improve predictive performance [67]. In this thesis, we have used a hybrid modeling approach wherein an unsupervised machine learning method of clustering is used to cluster the similar components together and then a supervised learning method is used to predict the best component from the cluster. For clustering algorithms such as Hierarchical agglomerative clustering, BIRCH, DBSCAN, OPTICS and Mean_Shift is used. The supervised ML techniques used are K-nearest neighbor, Random Forest, Support Vector machine, Gaussian naive bayes and XGBoost.

RQ2: To predict the best electronic component using machine learning method, which algorithms could prove to be the most efficient?

To predict the best electronic component to be used in design, machine learning techniques are used in two stages. In the first stage, using an unsupervised ML clustering method, clusters of similar and homogeneous electronic components are formed. In the next stage, using supervised ML learning, the best component in the cluster is recommended or predicted based on user query.

To discover the best clustering model, five clustering algorithms viz, Hierarchical agglomerative, DBSCAN, BIRCH, OPTICS and Mean shift are used. BIRCH and agglomerative are hierarchical methods, DBSCAN and OPTICS a density-based clustering method and Mean_Shift a kernel-based density function method. Metrics such as Silhouette score, Calinski-Harabasz score and Davies-Bouldin score are used to validate cluster results. In addition to these measures, the produced cluster outcomes are evaluated using business insights. The most appropriate result is then selected.

The validation index score for Agglomerative clustering were the highest and obtained clusters do match with expert insights. Based on the analysis of clustered results and expert insights, Hierarchical agglomerative clustering was yielding the best cluster result with the highest silhouette score amongst all other algorithms for both the dataset. Amongst all the supervised learning classification model used, K-Nearest

Neighbor classifier performs better with an accuracy score of 0.9531, precision score of 0.9505, Cohen_kappa score of 0.9529, Recall score of 0.9531 and f1 score of 0.9486. Hence, using a combination of Hierarchical agglomerative clustering and K-nearest neighbor ML technique the model is developed.

RQ3: Can the proposed automated component selection approach using Machine learning method perform better compared to the traditional manual component selection approach?

The proposed model was able to predict 84% of the component accurately for capacitor dataset and 81% accurately for resistor dataset. While analyzing the model output result, an important aspect of few components having wrong data pricing was also highlighted. The correction of this data will lead to more accurate predictions. But overall, in comparison, the model performance is quite good for a new novel approach. The model can help in finding the most optimal component in a very short duration compared to traditional methods which in return would save a lot of precious working time for the company. In terms of reducing process complexity and reducing the product delivery time, the proposed model helps in achieving it.

RQ4: Can the developed ML model be able to predict accurate substitutes for obsolete electronics components?

Another factor to consider when choosing a component is the component's life cycle, which is especially important in the case of semiconductor IC chips. If the original component manufacturer or product maker no longer produces them, they are termed outdated. The word "obsolete" implies "gone away". The OCMs provide item discontinuation notifications informing customers of their last-time-buy alternatives and dates [3]. In addition to other variables, product life cycle is directly dependent on the life of the electronic parts used to build it. The component's current 'availability' is significant, but until when the component will be in supply is more critical. Because of their poor component selection, the designs and products end up being expensive. Picking a component that is about to cease could be disadvantageous for organization, whereas picking the electronic component that is new to the inventory puts the component's endurance in jeopardy. To mitigate the component's unpredicted hazards, a list of alternate components is needed. The proposed model predicts the top 5 best components in a cluster ranked based on the specified business rules

enlisted by the company such as low price, high commonality, low lead time. If a specific component has been obsolete, the user can choose the next component from the recommended list of 5 components. In this way the proposed model successfully predicts a substitute for the obsolete component.

Main goal: Evaluate the value of combining unsupervised and supervised machine learning to predict the optimal electronic component.

Various combinations of supervised-unsupervised ML techniques are used in different domains to achieve an optimal result compared to the existing traditional method. Most of the studies gathered from the related works section have mentioned the significant benefit of combining supervised and unsupervised machine learning hybrid approach in their research. [17] have proposed a two-level hybrid approach comprising of unsupervised clustering methods and decision trees with boosting and evaluating the approaches based on the best results. [37] have also proposed a technique to improve the product classification accuracy in e- commerce domain by combining clustering with classification technique. Results show that applying clustering techniques prior to classification improves the accuracy of the classification model. [18] has proposed a hybrid classification framework based on clustering. In the first step, clustering algorithm is used to divide the dataset into N number of clusters. In the second step, clustering-based feature selection measure is built, i.e., the hybrid information gain ratio, thereafter, training C4.5 decision tree is performed using hybrid information gain ratio. The result showed a hybrid approach performing far better than the traditional supervised classification model.

In this thesis, the first task was to identify which machine learning technique could be applied to improvise and make electronic component selection process more efficient and faster. As the component of both datasets had no label using which only a supervised learning approach can be executed or by just grouping the similar components into clusters would not have solved our problem statement. Hence, it was needed to use a combination of supervised and unsupervised machine learning techniques. Through our research we were not able to find any previous work specifically executed to solve the electronic component selection process for component engineering domain. In total 5 clustering and 5 supervised learning methods were used and the best amongst those were selected to develop the model.

In this way, the supervised and unsupervised ML model with the best metric score for this dataset is selected. Hence, the approach of using a combination of unsupervised and supervised Machine Learning techniques proved to be quite efficient in predicting the best electronic component.

7.3 Research Limitations

Performance usually improves when more data becomes available, as it does with any machine learning technique. Having a larger number of component entries might improve results. This is also true while reviewing cluster outcomes, particularly since the business viewpoint was extremely helpful in this process. The greater the number of components in a cluster, the simpler it is to spot patterns in the cluster.

Missing data proved to be a major factor for the data exploration section. The dataset had missing data for very important attributes such as price, voltage, commonality in large numbers. As the goal of the study is to develop a model which will recommend an optional component with low price being one of the attributes, missing data for price attribute is quite challenging to impute.

According to our knowledge, very little literature study is available related to application of machine learning in the electronic component selection process. Hence, we had to look for other domain studies to relate with the component engineering domain to find relevant techniques.

We used data for capacitor and resistor in our study, although the approach and model can be used for other electronic components for the same business goal. Although few parameters need to be changed, the model technique will remain the same to generalize for other electronic components.

7.4 Future Work

We make several recommendations or proposals to other authors for further research studies based on our findings, the problems we experienced, and the results obtained. In future studies, one of the suggestions is to identify more factors or attributes such as (technical, electrical and environmental) affecting the component life cycle. This

helps in understanding the features to be selected for finding the substitutes for the obsolete components. In our study we mainly focused on technical and commercial parameters.

This study is limited to only two component datasets viz. capacitor and resistor but the model can be used for other electronic components also to predict the best one to use in design and manufacturing.

An ensemble approach would also be suggested for future work. In our study we have used a hybrid approach of combining supervised and unsupervised learning, but an ensemble learning technique can also be used for result improvement.

A work scope to use this study findings in demand forecasting of component and supplier selection could also be a challenging future work to be explored by researchers. In this way, companies can plan to deal with specific suppliers whose components are more used and can also negotiate with them for pricing.

References

[1] J. S, "The Principles and Process of Electronic Component Selection," Electronics Engineering Herald, 16 February 2017. [Online]. Available: http://www.eeherald.com/section/sourcing-database/component_sourcing_guide5.html.

[2] M. A. &. F. S. E. Waller, "Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management," *.Journal of Business Logistics,* vol. 34(2), pp. 77-84, 2013.

[3] J. S, "Electronic Component Sourcing: Supplier and Component Selection guide," Electronics Engineering Herald, 16 February 2017. [Online]. Available: http://www.eeherald.com/section/sourcing-database/component_sourcing_guide4.html.

[4] U. M. P.-S. G. &. S. P. Fayyad, "Knowledge Discovery and Data Mining: Towards a Unifying Framework," *KDD-96,* vol. 96, pp. 82-88, 1996, August.

[5] P. G. R. &. R. E. M. Espadinha-Cruz, " A review of data mining applications in semiconductor manufacturing," *Processes,,* pp. 9(2), 305., 2021.

[6] A. M. M. &. F. P. Jain, "Data clustering: a review," *ACM Comput. Surv,* vol. 31, pp. 264-323, 1999.

[7] A. K. &. D. R. C. Jain, " Algorithms for clustering data.," *Prentice-Hall, Inc.,* 1988.

[8] H. &. K. R. Frigui, "A Robust Competitive Clustering Algorithm With Applications in Computer Vision," *IEEE Trans. Pattern Anal. Mach. Intell,* vol. 21, pp. 450-465, 1999.

[9] G. &. E. C. Hamerly, "Learning the k in k-means," *Advances in neural information processing systems,* vol. 16, 2003.

[10] R. H. Turi, " Clustering-based colour image segmentation," Monash University, Melbourne, 2001.

[11] S. A. E. A. P. Omran Mahamed G. H., "Dynamic clustering using particle swarm optimization with application in image segmentation," *Pattern Analysis and Applications,* vol. 8, no. 4, p. 332, 2005.

[12] E. &. M. J. S. Stromatias, "Supervised learning in spiking neural networks with limited precision," *International Joint Conference on Neural Networks (IJCNN),* pp. 1-7, 2015.

[13] H. S. D. &. S. S. Wenzel, "In Artificial Intelligence and Digital Transformation in Supply Chain Management: Innovative Approaches for Supply Chains.A literature review on machine learning in supply chain management," *Proceedings of the Hamburg International Conference of Logistics (HICL),* vol. 27, pp. 413-441, 2019.

[14] S. B. Z. I. &. P. P. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering,* vol. 160(1), pp. 3-24, 2007.

[15] B. &. C. S. Kitchenham, "Guidelines for performing systematic literature reviews in software engineering," 2007.

[16] S. R. P. V. V. &. B. K. S. Gaddam, "K-Means+ ID3: A novel method for supervised anomaly detection by cascading K-Means clustering and ID3 decision tree learning methods," *IEEE transactions on knowledge and data engineering,,* vol. 19(3), pp. 345-354, 2007.

[17] I. &. C. X. (. Bose, "Hybrid models using unsupervised clustering for prediction of customer churn," *Journal of Organizational Computing and Electronic Commerce,* vol. 19(2), pp. 133-151, 2019.

[18] J. e. a. Xiao, "A hybrid classification framework based on clustering," *IEEE Transactions on Industrial Informatics,* pp. 2177-2188, 2019.

[19] G. H. J. M. J. &. H. L. Wang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering," *Expert systems with applications,* vol. 37(9, pp. 6225-6232, 2020.

[20] T. W. F. C. J. Y. Y. &. C. X. Ma, "A hybrid spectral clustering and deep neural network ensemble algorithm for intrusion detection in sensor networks," *Sensors,* vol. 16(10), p. 1701, 2016.

[21] R. Eslamloueyan, "Designing a hierarchical neural network based on fuzzy clustering for fault diagnosis of the Tennessee–Eastman process," *Applied soft computing,* pp. 1407-1415., 2011.

[22] A. M. I. &. M.-A. N. Heidarzade, "Supplier selection using a clustering method based on a new distance for interval type-2 fuzzy set sA case study.," *Applied Soft Computing,* vol. 38, pp. 213-231, 2016.

[23] R. &. M. J. (. Rajamohamed, "Improved credit card churn prediction based on rough clustering and supervised learning techniques," *Cluster Computing,* vol. 21(1), pp. 65-77, 2018.

[24] P. Kaur and A. Gosain, "Robust hybrid data-level sampling approach to handle imbalanced data during classification," *Robust hybrid data-level sampling approach to handle imbalanced data during classification,* vol. 24, 2020.

[25] X. Hu, J. Xu and J. Wu, "A Novel Electronic Component Classification Algorithm Based on Hierarchical Convolution Neural Network.," *IOP Conference Series: Earth and Environmental Science,* vol. 474., no. IOP Publishing, p. 5, 2020.

[26] A. M. I. &. M.-A. N. Heidarzade, "Supplier selection using a clustering method based on a new distance for interval type-2 fuzzy sets," *Applied Soft Computing,* vol. 38, pp. 213-231, 2016.

[27] S. B. K. &. R. M. Forouzandeh, "Presentation of a recommender system with ensemble learning and graph embedding: a case on MovieLens.," *Multimedia Tools and Applications,* vol. 80(5), pp. 7805-7832, 2021.

[28] C. V. N. &. S. A. Kaewchinporn, " A combination of decision tree learning and clustering for data classification," *Eighth international joint conference on computer science and software engineering (JCSSE),* pp. 363-367, 2011.

[29] I. M. F. E. M. F. F. A. &. I. D. Cavalcante, "A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing," *International Journal of Information Management,* vol. 49, pp. 86-97, 2019.

[30] K. R. &. V. C. M. Kashwan, "Customer segmentation using clustering and data mining techniques," *International Journal of Computer Theory and Engineering,* p. 856, 2013.

[31] C. F. L. C. W. &. C. S. C. Chien, "Analysing semiconductor manufacturing big data for root cause detection of excursion for yield enhancement," *International Journal of Production Research,* vol. 55(17), pp. 5095-5107, 2018.

[32] G. A. S. A. P. S. M. S. &. B. A. Susto, "Machine learning for predictive maintenance: A multiple classifier approach," *IEEE transactions on industrial informatics,* vol. 11(3), pp. 812-820, 2018.

[33] K. &. G. H. Gokcesu, "Natural Hierarchical Cluster Analysis by Nearest Neighbors with Near-Linear Time Complexity," *arXiv preprint arXiv:2203.08027.,* 2022.

[34] Y. K. &. S. K. Alapati, "Combining clustering with classification: a technique to improve classification accuracy," in *Lung Cancer*, 2018.

[35] C. F. &. L. Y. H. Tsai, "Customer churn prediction by hybrid neural networks.," *Expert Systems with Applications,* vol. 36(10), pp. 12547-12553, 2009.

[36] G. H. J. H. Y. L. J. M. M. S. Z. .. &. W. Y. Huang, "Machine learning for electronic design automation: A survey.," *ACM Transactions on Design Automation of Electronic Systems (TODAES),* vol. 26(5), pp. 1-46, 2021.

[37] N. M. N. G. N. A. M. &. J. R. M. Mathivanan, " Improving classification accuracy using clustering technique," *Bulletin of Electrical Engineering and Informatics,* 2018.

[38] D. Wu, " A hybrid model using DEA, decision tree and neural network," *Expert systems with Applications,* vol. 36(5), pp. 9105-9112, 2019.

[39] D. &. S. Stanisavljevic, " A Review of Related Work on Machine Learning in Semiconductor Manufacturing and Assembly Lines," 2016.

[40] G. L. Y. &. D. X. (. Zhang, "K-Means clustering-based electrical equipment identification for smart building application.," *Information,* Vols. 11(1), 27, 2020.

[41] V. A. J. &. L. C. P. Maxville, "Intelligent component selection," in *In Proceedings of the 28th Annual International Computer Software and Applications Conference*, 2004.

[42] Y. Y. G. L. J. &. H. J. Xu, "An electronic component recognition algorithm based on deep learning with a faster SqueezeNet," *Mathematical Problems in Engineering,* 2020.

[43] W. L. N. &. Y. K. Bao, " Integration of unsupervised and supervised machine learning algorithms for credit risk assessment," *Expert Systems with Applications,* vol. 128, pp. 301-315, 2019.

[44] Y. Y. G. L. J. &. H. J. Xu, "An electronic component recognition algorithm based on deep learning with a faster SqueezeNet.," *Mathematical Problems in Engineering,* 2020.

[45] R. &. H. J. Wirth, "CRISP-DM: Towards a standard process model for data mining," *In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining,* vol. 1, pp. 29-40, 2000.

[46] P. C. J. K. R. K. T. R. T. S. C. &. W. R. C. D. Chapman, "1.0: step-by-step data mining guide," p. 78, 2000.

[47] D. M. R. &. S. K. Patel, " A comparative study of clustering data mining: techniques and research challenges.," *International Journal of Latest Technology in Engineering, Management & Applied Science,* vol. 3(9), pp. 67-70., 2014.

[48] N. Mlambo, "Data Mining: Techniques, Key Challenges and Approaches for Improvement," *International Journal of Advanced Research in Computer Science and Software Engineering,,* vol. 6(3), pp. 59-65., 2016.

[49] J. I. I. O. F. A. O. U. E. A. F. &. A. E. Oyelade, "Clustering algorithms: their application to gene expression data.," *Bioinformatics and Biology insights,* 2016.

[50] J. C. S. S. V. &. M. R. Lee, " Service cost estimation for packaged business application-based business transformation," *In 2008 IEEE International Conference on Service Operations and Logistics, and Informatics ,* vol. 1, 2008.

[51] J. Handl, "Computational cluster validation in post-genomic data analysis," *Bioinformatics,* vol. 21, no. 15, pp. 3201-3212, 2005.

[52] M. K. H. P. S. J. &. X. X. Ester, " A density-based algorithm for discovering clusters in large spatial databases with noise," vol. 96, pp. 226-231, 1996.

[53] M. B. M. M. K. H. P. &. S. J. Ankerst, "OPTICS: Ordering points to identify the clustering structure," *ACM Sigmod record,* vol. 28(2), pp. 49-60, 1999.

[54] P. H. J. F. M. &. M. J. Trebuňa, " The importance of normalization and standardization in the process of clustering," *In 2014 IEEE 12th International Symposium on Applied Machine Intelligence and Informatics ,* 2014.

[55] G. W. &. C. M. C. Milligan, "Methodology review: Clustering methods," *Applied psychological measurement,* vol. 11(4), pp. 329-354, 1987.

[56] K. A. &. T. P. Patel, "The best clustering algorithms in data mining," *In 2016 International Conference on Communication and Signal Processing (ICCSP) IEEE,* pp. 2042-2046, 2016.

[57] D. B. (. Rubin, " Inference and missing data," *Biometrika,* vol. 63(3), pp. 581-592, 1976.

[58] "Secondary analysis of electronic health records," in *Critical Data, M. I. T.*, Springer Nature, 2016, p. 427.

[59] D. &. T. Y. (. .. ,. Xu, "A comprehensive survey of clustering algorithms," *Annals of Data Science,* vol. 2(2), pp. 165-193, 2015.

[60] V. &. Y. J. Estivill-Castro, "Cluster validity using support vector machines," *In International Conference on Data Warehousing and Knowledge Discovery,* pp. 244-256, 2003.

[61] J. K. J. &. K. D. B. Handl, "Computational cluster validation in post-genomic data analysis," *Bioinformatics,* vol. 21(15), pp. 3201-3212, 2005.

[62] A. Hardy, "On the number of clusters," *Computational Statistics & Data Analysis,* vol. 23(1), pp. 83-96, 1996.

[63] O. G. I. M. J. P. J. M. &. P. I. Arbelaitz, "An extensive comparative study of cluster validity indices," *Pattern recognition,* vol. 46(1), pp. 243-256, 2013.

[64] J. Ansaharju, "Improving Software Development with Platform-as-a-Service Product– Using Heroku in Web Application Project.," 2016.

[65] V. A. J. &. L. C. P. Maxville, "Intelligent component selection," *In Proceedings of the 28th Annual International Computer Software and Applications Conference, 2004. COMPSAC. IEEE.,* pp. 244-249, 2004.

[66] I. M. F. E. M. F. F. A. &. I. D. Cavalcante, "A supervised machine learning approach to data-driven simulation of resilient supplier selection in digital manufacturing.," *International Journal of Information Management,* pp. 49, 86-9, 2019.

[67] Y. K. &. S. K. Alapati, " Combining clustering with classification: a technique to improve classification accuracy.," *Lung Cancer,* pp. 32(57), 3., 2016.

[68] S. K. S. A. M. &. P. R. Bodare, "Crime Analysis using Data Mining and Data Analytics," *International Research Journal of Engineering and Technology,* pp. 7679-7682, 2019.

[69] A. K. M. N. M. a. P. J. F. Jain, "Data Clustering: A Review," *ACM Computing Surveys,* vol. 31, p. 264–323, September 1999.

[70] Z. a. Karypis, "Evaluation of Hierarchical Clustering Algorithms for Document Datasets," *Proceedings of CIKM,* 2002.

[71] M. G. K. a. V. K. Steinbach, "A Comparison of Document Clustering Techniques," *KDD Workshop on Text Mining,* 1999.

[72] A. K. M. M. N. &. F. P. J. Jain, "Data clustering: a review," *ACM computing surveys (CSUR),* vol. 31, pp. 264-323, 1999.

[73] Y. &. K. G. Zhao, "Evaluation of hierarchical clustering algorithms for document datasets," *In Proceedings of the eleventh international conference on Information and knowledge management,* pp. 515-524, 2002, November.

[74] M. K. G. &. K. V. Steinbach, "A comparison of document clustering techniques.," 2000.

[75] C. v. M. M. G. A. G. L. V. C. N. T. &. F. A. Krahé, "Sensitivity to CT-optimal, affective touch depends on adult attachment style," *Scientific reports,* vol. 8(1), pp. 1-10, 2018.

[76] I. H. F. E. H. M. A. P. C. J. Witten, "Practical machine learning tools and techniques. In DATA MINING," vol. 2, p. 4, 2005.

[77] M. &. H. N. R. Gusenbauer, "Which academic search systems are suitable for systematic reviews or meta-analyses?," *Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. Research synthesis methods,* vol. 11(2), pp. 181-217, 2020.

[78] P. N. S. M. &. K. V. Tan, "Data mining introduction. People's Posts and Telecommunications," *Publishing House, Beijing,* 2006.

[79] F. V. G. G. A. M. V. T. B. G. O. .. &. D. E. Pedregosa, " Scikit-learn: Machine learning in Python," *The Journal of machine Learning research,* vol. 12, pp. 2825-2830, 2011.

[80] I. B. &. U. D. Mohamad, "Standardization and its effects on K-means clustering algorithm.," *Research Journal of Applied Sciences, Engineering and Technology,* vol. 6(17), pp. 3299-3303, 2013.

[81] R. O. &. H. P. E. Duda, Pattern classification and scene analysis, New York: Wiley, 1973.

[82] A. J. Z. A. G. M. H. M. S. H. F. Z. A. .. &. K. A. P. Abadi, "Small in size, but large in action: microRNAs as potential modulators of PTEN in breast and lung cancer," *Biomolecules,* vol. 11(2), p. 304, 2021.