공학석사 학위논문

Data Mining Method for Offshore Structures based on Big Data Technology

빅데이터 기술을 이용한 해양구조물의
데이터 마이닝 방법

2019년 2월

서울대학교 대학원
협동과정 해양플랜트엔지니어링 전공
박 성 우

Contents

Figures

Tables

Abstract

Data Mining Method for Offshore Structure based on Big Data Technology

Sung-Woo Park
Interdisciplinary Program in Offshore Plane Engineering
The Graduate School
Seoul National University

As many products as ships and offshore structures are constructed in the shipyard, and various data are generated and stored in the design or construction stage. Big data technology needs to be applied to process data of large size quickly, obtain meaningful results and use it for decision making. In this paper, we propose a solution to two of the problems that may occur in the shipyard.

One of the two problems which can arise in the shipyard has mainly happened in the design stage. Engineers can make the mistake of choosing the wrong material in the design process, and the wrong material selection in the design process can directly lead to a design error. Another problem may arise during the procurement and purchase process. In the absence of additional information such as lead time of material or inventory at the time of procurement, additional time is required to retrieve the data. Both problems arise predominantly from the unskilled. Therefore, the purpose of this study is to establish a system that can inform the engineers about the relationships between materials which can be obtained by association analysis and material requirements which can be obtained by regression analysis. This kind of system can help the engineers to reduce design errors and time consuming due to the procurement process.

The information of piping materials used in an offshore structure can be regarded as ‘big data’ because of their various types and size, and the data mining algorithms based on the big data technology are applied to data related to the offshore structures. To analyze the relationship between materials for design, ‘frequent pattern growth algorithm’ was used. For material requirement analysis, big data technology-based regression analysis was used to generate a regression model, respectively.

Finally, the proposed method was used to check the relationship between materials, and to predict material requirement, and verified the effectiveness of the proposed method by comparing each result with actual cases.

Keywords: Big data, Data mining, Offshore structure, piping design, Association analysis, Regression analysis.

Student number: 2017-20003

1. Introduction

1.1. Research background

(1) Necessity of big data technology

When manufacturing products in the shipbuilding and offshore structure industries, it takes two main steps: engineering and construction. Once the construction is completed, it is delivered to the operators and the operational phase begins. At each step and phase, various types of data are produced, and the shipyards are making effort to store and process the generated data for use in decision making. In the case the Korean shipyards, in the meantime, if data produced by each discipline are separately collected and processed, and the data which is necessary for decision making are made. Now a day, there is a need to automatically create the data which is necessary for decision making form the database. This is due to the problem of data reliability and processing speed. First, there is possibility that the data will be distorted in favor of the discipline. Second, depending on organization, flow of data; collecting, processing, and reporting can take a long time. In order to solve problems that arise because people directly deal with data, there is a need to process data using computer technology such as big data or data mining.

The shipyards use a variety of information during decision making. In terms of engineering, they use engineering progress, ratio of error, completeness of engineering. In another hand, for procurement, they use lead time of equipment or material, stock management, material requirement, progress of project for decision making.

Figure 1 is briefly describing data produced from shipbuilding and offshore industry and how it is processed and used for decision making.

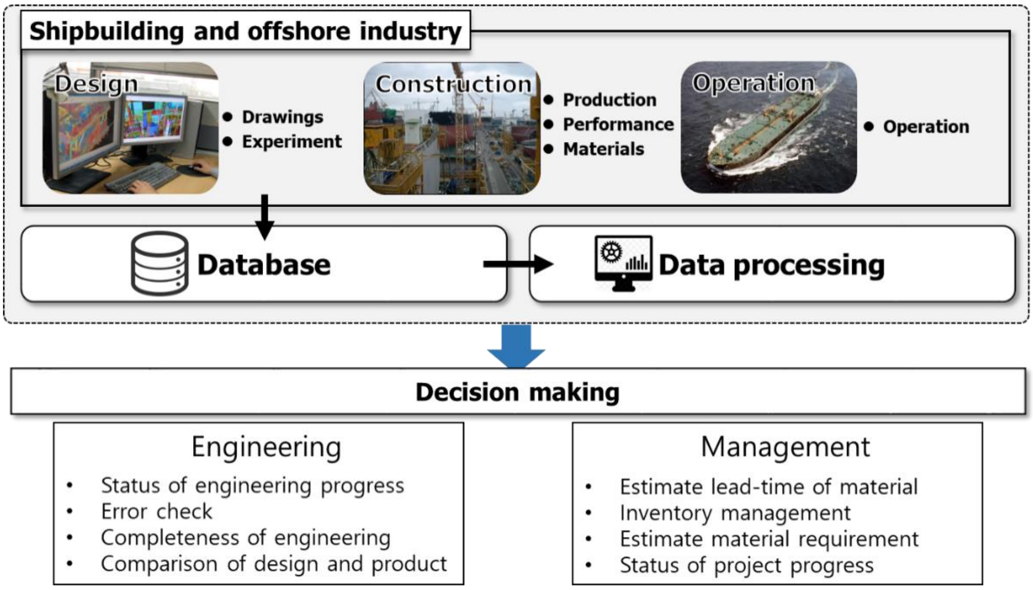

Shipbuilding and offshore industry
Design
Drawings
Experiment
Construction
Production
Performance
Materials
Operation
Operation
Database
Data processing
Decision making
Engineering
Status of engineering progress
Error check
Completeness of engineering
Comparison of design and product
Management
Estimate lead-time of material
Inventory management
Estimate material requirement
Status of project progress

Figure 1 Data produced from shipyard

Table 1 is comparing the data processing technology; traditional data processing and big data technology. Traditional data processing technology has some disadvantage. First one is slow processing time. There is limitation of processing unit in traditional data processing, and it cause bottleneck for the data processing. Second one is limitation of memory. The size of data is getting bigger and bigger. So, it requires a lot of memory to process such a big data. The last one is that it required post-processing step to show result of analysis to make people understand.

The big data technology is different. Due to its characteristics; distributed processing, it shows faster processing time compared with traditional data processing. When we use big data technology, each cluster is sharing their processing unit and memory. So, we don't have limitation of memory, too. This characteristic makes us possible to stream large size of data in real time. In terms of visualization, big data technology has interactive visualization function and it makes visualization easier compared with traditional technology. This study aims to apply these advantages of big data technology to the shipbuilding and offshore industry, especially for offshore structure.

Table 1 Comparison of data processing between traditional technology and big data technology

Traditional data processing	Big data technology
- Slow processing time - Limitation of memory - Post-processing for visualization	- Fast processing by distributed processing - Large data streaming and processing - Easy visualization

(2) Things to consider during engineering and material procurement of offshore structures

In this chapter, we will look at the appropriate problems to apply Big Data technology in two steps in building offshore structures, engineering and procurement. To construct an offshore structure, these two steps are very important. If the engineering is not done properly, the correct production cannot be constructed. And if there is any problem with procurement, construction will be delayed and not be done on time. Due to these reasons, many things regarding engineering and procurement must be considered during manufacturing offshore structure. is describing problem from engineering and procurement process of an offshore structure and things to consider. In engineering, materials meet project specification, materials meet the given pressure or temperature should be considered. In addition, piping support that can make pipeline overcome its weight should be included into the consideration. In terms of procurement, timing of purchase of many materials, lead time, materials in stock should be considered. Therefore, the engineer or the purchasing manager needs a lot of experience in order to proceed all the processes smoothly. It takes additional time to search for the possibility of error in engineering and material information to buy while observing various conditions. Figure 2 is summarizing these problems

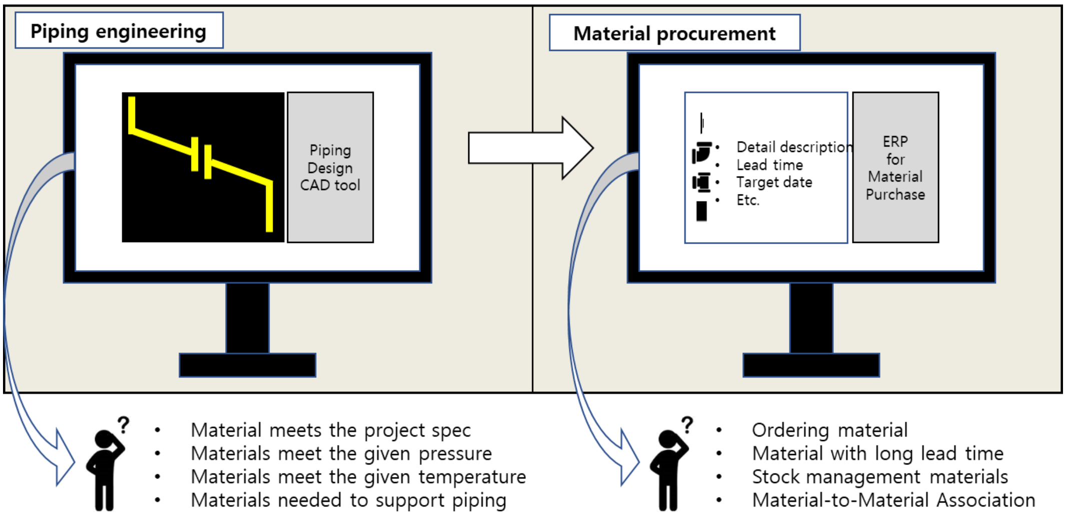

Piping engineering
Piping Design CAD tool
Material procurement
Detail description
Lead time
Target date
Etc.
ERP for Material Purchase
?
Material meets the project spec
Materials meet the given pressure
Materials meet the given temperature
Materials needed to support piping
?
Ordering material
Material with long lead time
Stock management materials
Material-to-Material Association

Figure 2 Problem occur during piping engineering and material procurement

These problems can usually be solved by many experiences. Experienced engineers and procurement personnel are already aware of problems that may arise during the engineering or purchase process, and in case of problems, they may already know the solutions or can find the solutions easily.

Table 2 Related works

Related Works	Year	Application	Data technology	Data Mining	Remark
Li et al.	2015	Big Data in product lifecycle management	-	-	the existing applications of "Big Data" in PLM are summarized
Tapedia et al.	2016	Data mining for various Internet of Things applications	RDBMS	-	No application
Ham et al.	2016a	Procurement management of shipyard equipment	Azure	Multidimensional regression Boosted decision tree regression, Neural network regression	
Ham et al.	2016b	Machine Learning Application for using Shipyard Big Data	Azure	Decision tree regression	
Ham	2016	Offshore plant's outfitting procurement management	RDBMS	Multidimensional regression	
Saabith et al.	2016	Apriori algorithms on the Hadoop MapReduce platform	Hadoop	Association analysis (Apriori algorithm)	
Zhang et al.	2017	Product lifecycle management	Proposed framework	-	No application
Li et al.	2017	Smart manufacturing	Hadoop	Proposed algorithm	
Lee	2017	Reference Model for Big Data Analysis in Shipbuilding Industry	-	-	No application
Kim et al.	2017	Weight estimation of FPSO topside	Big data (Hadoop)	Multidimensional regression	
Musalem et al.	2018	Market basket analysis insights to support category management	-	multidimensional scaling and clustering	
Changhai et al.	2018	Factors correlation mining on maritime accidents	RDBMS	Association analysis (Apriori algorithm)	
Abbasian et al.	2018	Improving early OSV design robustness	Big data (Hadoop)	Clustering	
Griva et al.	2018	Analyzing customer visit	-	Clustering	
He et al.	2018	Impact of urban growth pattern	-	Association analysis (Apriori algorithm)	
Park et al.	2018	Text mining for transportation management plan	-	Association analysis (Apriori algorithm)	
Szymkowiak et al.	2018	Applying market basket analysis to official statistical data	-	Association analysis (Apriori algorithm)	
Oh et al.	2018	Estimation of material requirement for offshore structure	Big data (Hadoop)	Multidimensional regression	
This Study	2018	Data mining of data from offshore structure based on big data technology	Big data (Hadoop)	Association analysis (Frequent patterns growth algorithm) Multidimensional regression	

1.2. Related works

It is examined related works to see if there were similar problems. Table 2 is listing related works. In various industries, big data or data mining technology is used to solve the problems of each industry. However, there have been very few application researches that applied big data technology to real problems in the shipbuilding and offshore industry. is summarizing related work in terms of their application, data processing technology and data mining algorithm which they used. There are two categories in related works. One is that research focus on data mining mythology to solve a specific problem in an industry. Another one is that adopts big data technology to find meaningful data from unclassified data.

First category include study such as Li et al. (2015) which summarized the existing applications of "Big Data" in product lifecycle management. Tapedia et al. (2016) researched on data mining for various Internet of Things application. Ham et al. (2016) is related to machine learning application for using big data from shipyard. Ham (2016) is another research regarding procurement management of shipyard. This study is for methodology of data mining algorithm based on multidimensional regression. Zhang et al. (2017) proposed a framework for product lifecycle management. Musalem et al. (2018) [supports category management using market basket analysis insights. In terms of maritime, Changhai et al. (2018) suggests factors correlation mining on maritime accidents. This study uses association analysis. Griva et al. (2018) used clustering for analyzing customer visit.

Also, following three studies are using association analysis for their main algorithm. He et al. (2018) is about impact of urban growth pattern. Park et al. (2018) studies text mining for transportation management plan. Szymkowiak et al. (2018) applies market basket analysis to official statistical data.

Ham et al. (2016a) is about procurement management of shipyard. This study is based on big data technology. Saabith et al. (2016) studies apriori algorithms on the Hadoop MapReduce platform which is one of main big data technology. Li et al. (2017) proposed an algorithm based on Hadoop ecosystem for smart manufacturing. Lee (2017) propose a reference model for big data analysis in shipbuilding industry.

Kim et al. (2017) is about weight estimation of FPSO (Floating, Production, Storage and Offloading) using multidimensional regression based on Hadoop ecosystem. In terms of design, Abbasian et al. (2018) suggests how to improve early OSV (Offshore Support Vessel) design using clustering on Hadoop. At last, Oh et al. (2018) estimates material requirement for offshore structure based on Hadoop.

The purpose of this study is to study data mining applications based on Big Data Technology for information on offshore structures.

1.3. Target of the study

Looking back at the problems which are mentioned in the previous chapter, there are a lot of things to consider in the engineering and procurement process, and user experience is required to prevent any kind of engineering error and to save time. Without enough experience, there is significant potential for engineering error and additional time required to search material information. Figure 3 is summarizing proposal of this study to solve the problems.

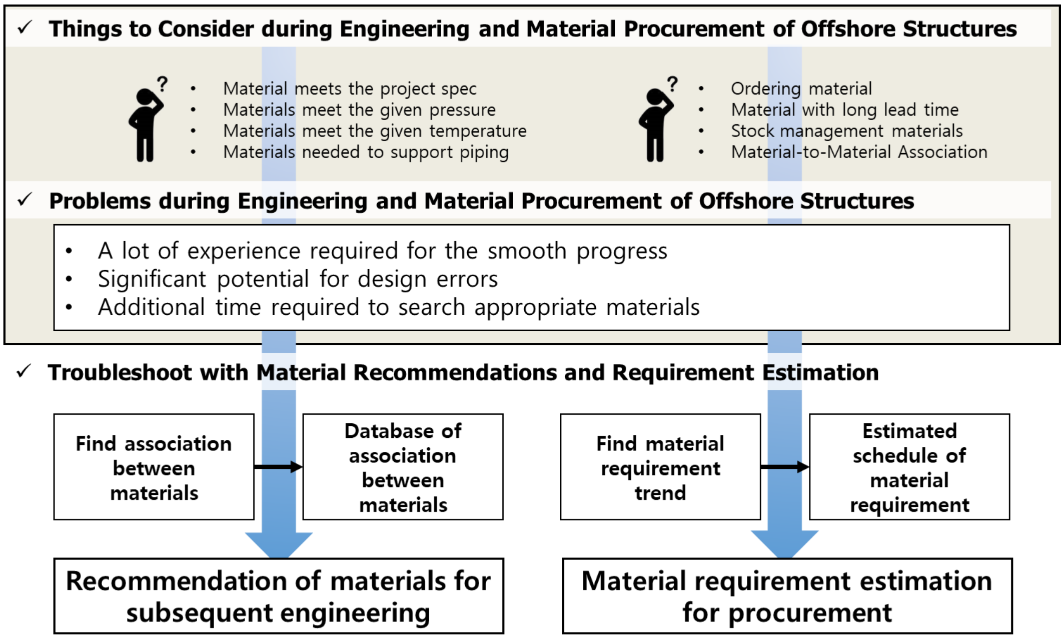

✓ Things to Consider during Engineering and Material Procurement of Offshore Structures
?
• Material meets the project spec
• Materials meet the given pressure
• Materials meet the given temperature
• Materials needed to support piping
?
• Ordering material
• Material with long lead time
• Stock management materials
• Material-to-Material Association
✓ Problems during Engineering and Material Procurement of Offshore Structures
• A lot of experience required for the smooth progress
• Significant potential for design errors
• Additional time required to search appropriate materials
✓ Troubleshoot with Material Recommendations and Requirement Estimation
Find association between materials
Database of association between materials
Find material requirement trend
Estimated schedule of material requirement
Recommendation of materials for subsequent engineering
Material requirement estimation for procurement

Figure 3 Target of study

In this study, we propose providing information through data mining based on big data technology as solution of the problem. In the case of engineering, we use association analysis between various materials to recommend materials suitable for the engineering. For procurement process, we will make a way to save time by analyzing and predicting the required amount of material needed at each point in time. In the other words, by data mining the data related to offshore structure and extracting the knowledge that is helpful for engineering and procurement, we aim to create a basis system that supports the scarce experience.

1.4. System configuration

Figure 4 shows system configuration of this study. First, there is input module which is to understand information related to offshore structure, especially its materials. There are several types of information such as type of material, material selection, schedule & rating and data from 3D CAD (Computer aided design) tools.

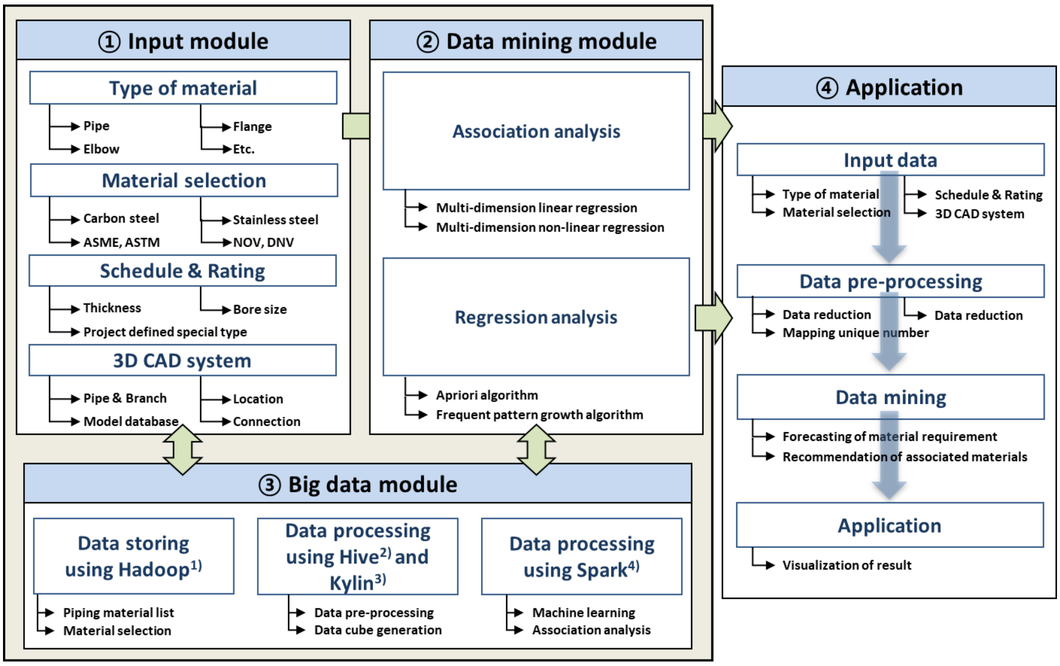


Figure 4 System configuration

We also have data mining module. It includes two main function; association analysis and regression analysis. The association analysis is to find relationships between materials and return the information to users, so the users can find appropriate materials during their engineering job. Regression analysis part trains a regression model from material requirement of offshore structures. Trained regression model is used to predict material requirement of new offshore structure. Big data module is basis of two modules. It is including data storage unit and data processing unit. In this study, Hadoop and its ecosystem are used as big data technology and details will be described later. And these three parts consist the big data framework.

The last one is application. In this application, all actions are taken to process and analyze the data such as getting input data, pre-processing, data mining and visualization. As user check the visualized result from the application, they can see the data to use in decision making, including engineering information such as material association.

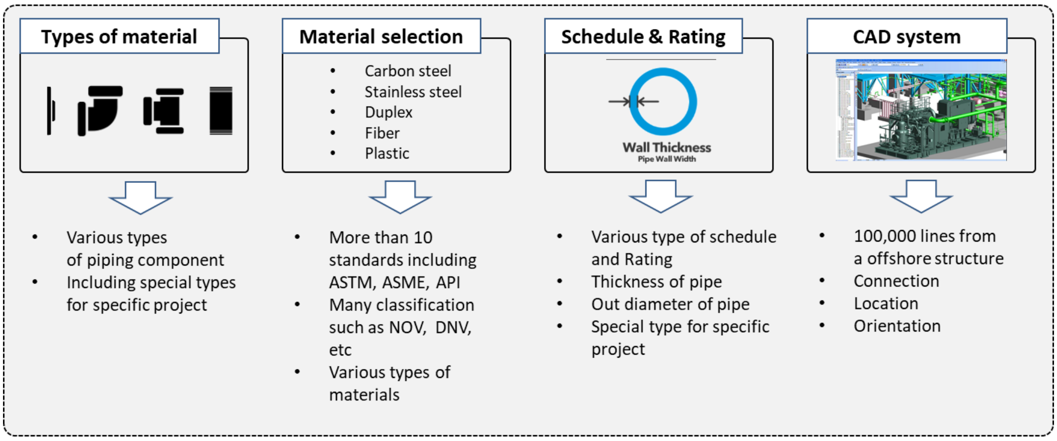
2. Data mining method for pipng material

This chapter describes data mining methods using piping materials. First, the reason why the piping material is used as the source of the big data will be explained, and how to solve the problems that may occur during the construction process of the offshore structure by using the piping materials will be explained. We used association analysis and regression analysis to solve the two problems mentioned above. Association analysis can be used as a basis of system that identify material relationship and recommend materials that are likely to be used with the materials used by engineers during engineering phase. Regression analysis aims to learn the requirements of the material during construction of the offshore structures and then to predict the material requirements of the next project so that it can help to make good purchase plan.

.

2.1. Piping materials of an offshore structure

Piping related data is diverse, and the amount of data accumulated in various offshore structures is enormous. The cumulative data can be advantageously speeded up by applying big data technology rather than traditional data processing methods. Figure 5 is about the information belonging to the piping materials used in offshore structures. Since the piping materials are used not only with piping but also with various materials such as flanges, elbows and tees, the combinations are very diverse. In addition, different materials should be used for each system according to the design and operation specifications of offshore structures. Materials such as FRP and plastic as well as metal materials such as carbon steel and stainless steel may be used, and each material should meet the standards provided by ASTM and ASME. Likewise, it should have the thickness and diameter to fit the specification. Based on this information, the designer enters the information into the 3D cad system and starts the piping design work. As the piping design progresses, each material contains location information, orientation information, and connection information, which makes the information very complicated.



Types of material

- Various types of piping component
- Including special types for specific project

Material selection

- Carbon steel
- Stainless steel
- Duplex
- Fiber
- Plastic

- More than 10 standards including ASTM, ASME, API
- Many classification such as NOV, DNV, etc
- Various types of materials

Schedule & Rating

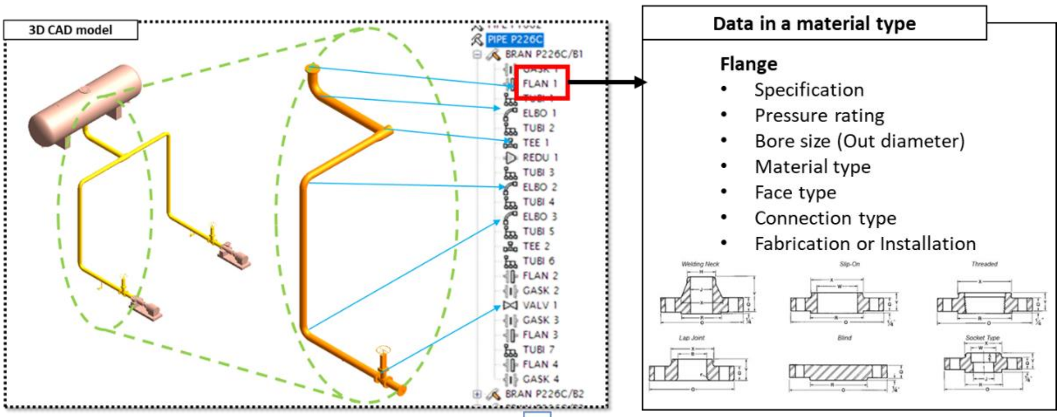
- Various type of schedule and Rating
- Thickness of pipe
- Out diameter of pipe
- Special type for specific project

CAD system

- 100,000 lines from a offshore structure
- Connection
- Location
- Orientation

Figure 5 Data from piping information

Figure 6 shows an example where piping information is combined. A pipe consists of several branches, each branch being listed in order of the materials used, as shown in the figure. Considering the flange as one of the starting materials, one flange contains information about the specifications, pressure, size, material, and connection type of the material, and additionally, the connection information. Depending on the connection information, whether the pipe is manufactured and installed, and whether the flange is installed directly on the site, additional information will be generated, and the information will be listed and managed.



Data in a material type

Flange

- Specification
- Pressure rating
- Bore size (Out diameter)
- Material type
- Face type
- Connection type
- Fabrication or Installation

List of piping material

Module	Pipe	Type in CAD	Type1	Material	Pressure Rating	Bore size (OD)(in)	Detail description
Module A	Pipe 1	FLAN	Flange	GRE	150	6	Heavy duty flange, Flat face, GRVE, #150, 6", NORSOK certified
Module A	Pipe 2	FLAN	Flange	GRE	150	2	Heavy duty flange, Flat face, GRVE, #150, 1.5", NORSOK certified
Module A	Pipe 3	FLAN	Flange	GRE	150	2	Heavy duty flange, Flat face, GRVE, #150, 1.5", NORSOK certified
Module A	Pipe 4	FLAN	Flange	GRE	150	2	Heavy duty flange, Flat face, GRVE, #150, 1.5", NORSOK certified
Module A	Pipe 5	FLAN	Flange	GRE	150	10	Heavy duty flange, Flat face, GRVE, #150, 10", NORSOK certified
Module A	Pipe 6	FLAN	Flange	GRE	150	10	Heavy duty flange, Flat face, GRVE, #150, 10", NORSOK certified

Figure 6 Example of data produce from piping component

You can measure the size of the materials used in a single offshore structure by using statistical analysis. Figure 7 is a statistic for the major materials used in an offshore structure. As can be seen from the figure, various materials are used extensively. In the case of elbow, 3 5549 pieces of 337 kinds are used, and in the case of the next most commonly used flanges, 17323 pieces of 348 kinds are used.

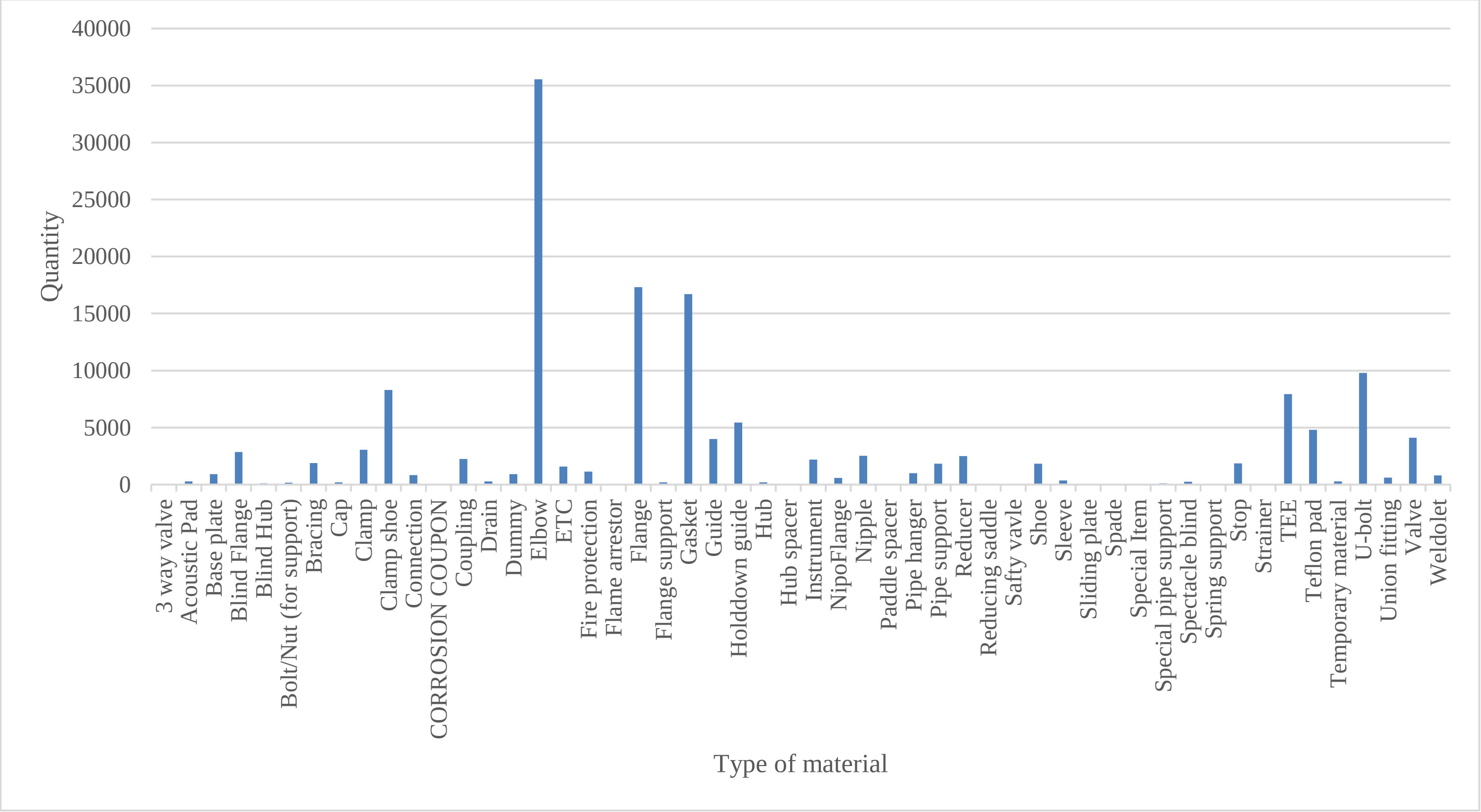


Figure 7 Statistics of piping components

2.2. Association analysis

In this chapter, we will explore data mining methods called association analysis and how association analysis can be applied to data mining of offshore structures.

2.2.1. Assotication between piping material for recommendation of associated materials

In the case of association analysis, it is a data mining algorithm that is mainly used for market basket analysis. We will examine how the algorithm works and how it can be applied to the association analysis of pipe materials and the material recommendation algorithm.

(1) Market basket analysis

The market basket analysis is one of the data analysis methods that are often used in retail data analysis. After consumers shop at the supermarkets or marts, they track what items are in the shopping cart and analyze the items that appear together to provide various insights. The shopping cart analysis mainly analyzes the items included in one shopping cart, the items purchased together, and finds out how the items are related to the buyer. Also, it can find items that are likely to be purchased together but are missing. These are summarized in Figure 8. Similar rules can be applied to offshore structures. Assuming a single shopping cart is a single pipe, each item contained in the shopping cart can be regarded as a pipe material used to construct the pipe. This will allow you to track which items are used together in a single piping, and to make recommendations for items that will appear but not. Furthermore, it is possible to compare the characteristics of the offshore structure and the offshore structure that frequently occur with specific combinations and find out what correlation there is.

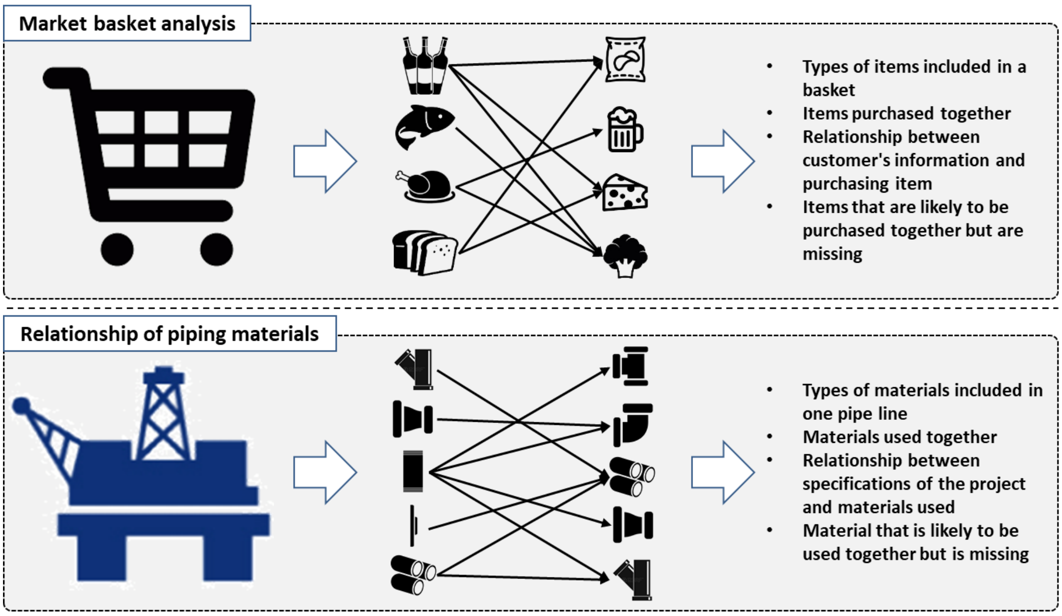

Market basket analysis
Types of items included in a basket
Items purchased together
Relationship between customer's information and purchasing item
Items that are likely to be purchased together but are missing
Relationship of piping materials
Types of materials included in one pipe line
Materials used together
Relationship between specifications of the project and materials used
Material that is likely to be used together but is missing

Figure 8 Market basket analysis of piping materials

(2) Principal of association analysis

Affinity analysis works in the following order: Assume first that there is a set of items like Table 3. First,

a. Check each itemset and count total amount of each itemset.

b. Check the support which is the probability that an item contains in an itemset.

c. Check the confidence which is conditional probability that an itemset having an item with another one

d. Find all the rules that X → Y with minimum support and minimum confidence that given by user

In this case, when given minimum support is 50% and minimum confidence is 50%, frequent pattern is itemset of beer and diaper with support 3.

Table 3 Example itemset for market basket analysis

Id	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	Nuts, Coffee, Diaper, Eggs, Milk

2.2.2. Comparison of association analysis algorithms

(1) Camparison of pattern mining algorithms and association mining algorithms

There are various algorithms that can perform association analysis. In this case, we must apply simple pattern mining algorithm and association analysis algorithm separately. The pattern mining algorithm aims to pick out items with a high frequency of occurrence in several item sets. The result is a set of items which appears most or more than the frequency specified by the user. However, in the case of association analysis, there is a difference in that the relationship between the items can be analyzed together because the conditional probabilities of specific items are calculated and presented together. Typical pattern mining algorithms and association analysis algorithms are summarized in Table 4.

Representative association analysis algorithms include apriori, frequent pattern growth algorithm, and top-k non-redundant association rule algorithm. In apriori algorithm, the pattern is analyzed, and the association is extracted by using array. The other two algorithms analyze the pattern with the tree type data structure supported by each algorithm and extract the association. Especially, in the case of FP-growth algorithm and top-k non-redundant association rule algorithm, unlike apriori, it is advantageous in that the processing speed is much faster because the pattern candidate is not generated while reading the database every time. Pattern mining has a traditional ECLAT, and recently algorithms such as PrePost and FIN have been developed to speed up.

The purpose of this study is to find the relation between the piping material items and to recommend the material to the designer through it. Therefore, we do not use the pattern mining algorithm that finds the simple combination of items, but association analysis algorithm was used.

In the next section, we show the process of determining which algorithm is used in this study through comparison of results and performance between association analysis algorithms.

Table 4 Comparison of association analysis algorithms and pattern mining algorithms

	Association analysis			**Frequent pattern mining**		
Name	**Apriori**	**FP-Growth**	**Top-K Non-redundant association rules**	**ECLAT**	**PrePost+**	**FIN**
Technique	Breadth first search & Apriori property (for pruning)	Divide and conquer	define top-K rather than confidence	Depth first search & intersection of transactions to generate candidate itemset	PPC-tree structure: N-list Pruning strategy: superset equivalence	Construct the POC-tree and identify all subsequent frequent item sets
Database scan	Each time a candidate item set generated	2 times only	2 times only	Few times (best case = 2)	Only 1 time	Only 1 time
Time	Execution time is considerable as time is consumed in scanning database	Less than Apriori algorithm	Similar with FP growth algorithm	Less than Apriori algorithm	Less than FP growth algorithm	Less than FP growth algorithm
Data format	Horizontal	Horizontal	Horizontal	Vertical	Vertical	Horizontal
Storage structure	Array	Tree (FP tree)	Tree	Array	N-list	Pre-Order Coding (POC_ tree)
Drawback	Too many candidate itemset Requires large memory space	FP-tree is expensive to build Consumes more memory	depends on redundancy **output is different by input 'k'**	Required virtual memory	FP-growth becomes more efficient and is faster when minimum support is small PrePost+ consumes a bit more memory than FP-growth	when the minimum support becomes small, the runtime of FIN is lesser compared to Prepost
Advantage	Use large itemset property. Easy to implement	No candidate generation	No candidate generation	No need to scan database each time a candidate itemset is generated	PrePost+ performs best compared to FP-growth	FIN run faster than PrePost and FP growth overall
Runtime	134ms	**26ms**	**20ms**	111ms	32ms	49ms

(2) Comparison of association algorithms in terms of results

In this chapter, we examine the results of the association analysis algorithm. First, we examined the results of the association analysis using Table 5 as the example.

Table 5 Example itemset

Id	Items bought
1	Beer, Nuts, Diaper
2	Beer, Coffee, Diaper
3	Beer, Diaper, Eggs
4	Nuts, Eggs, Milk
5	Nuts, Coffee, Diaper, Eggs, Milk

Table 6 is the result of the apriori algorithm. Support is the result of the ratio, the strongest combination being beer → diaper combination. This can be confirmed by confidence, which means that if the confidence is 1, the combination appears for every number of cases.

Table 6 Result of apriori algorithm

Item1	Item2	Support	Confidence
Beer	Diaper	0.5	1
Diaper	Beer	0.5	0.75
Coffee	Diaper	0.33333333	0.66666667
Diaper	Coffee	0.33333333	0.5
Coffee	Milk	0.33333333	0.66666667
Milk	Coffee	0.33333333	0.66666667
Diaper	Eggs	0.33333333	0.5
Eggs	Diaper	0.33333333	0.66666667
Diaper	Nuts	0.33333333	0.5
Nuts	Diaper	0.33333333	0.66666667

The following (Table 7) is the analysis result of the top-k non-redundant algorithm. In the case of top-k non-redundant algorithm, the goal is to find the combination of the user's input K value. It is important to note that support is represented by counting all occurrences, and combinations that do not appear depending on the value of K may also occur. Similarly, beer → diaper combination was the most frequent.

Table 7 Result of top-k non-redundant algorithm

Sets	Support	Confidence
Milk ==> Nuts	SUP: 2	1
Coffee ==> Diaper	SUP: 2	1
Milk ==> Eggs	SUP: 2	1
Milk ==> Nuts Eggs	SUP: 2	1
Eggs Milk ==> Nuts	SUP: 2	1
Nuts Eggs ==> Milk	SUP: 2	1
Nuts Milk ==> Eggs	SUP: 2	1
Beer ==> Diaper	SUP: 3	1

Finally, we examine the results of the frequent pattern growth algorithm (Table 8). In the case of this algorithm, the user will set the minimum support, and the frequencies below it will be truncated. The analysis is performed on several item sets, and the result is the beer → diaper combination like the other algorithms.

Table 8 Result of FP-growth algorithm

Item1	Item2	Confidence
Beer	Diaper	1
Coffee	Milk	0.666666667
Diaper	Eggs	0.5
Eggs, Milk	Nuts	1
Eggs, Nuts	Milk	1
Milk, Nuts	Eggs	1
Nuts	Eggs, Milk	0.666666667

As we have seen, all algorithms have the same results in terms of finding associations between items, although there are slightly different parts, such as how to handle support.

(3) Comparion of association algorithms in terms of runtime

Performance can be confirmed by comparing the execution time of each algorithm (Figure 9). The graph compares the execution times for the same piping material dataset. Compared with the FP-growth algorithm, the top-k algorithm shows better runtime performance. The top-k algorithm may have different associations in the results depending on the k values. So, in this study, FP-growth algorithm was used

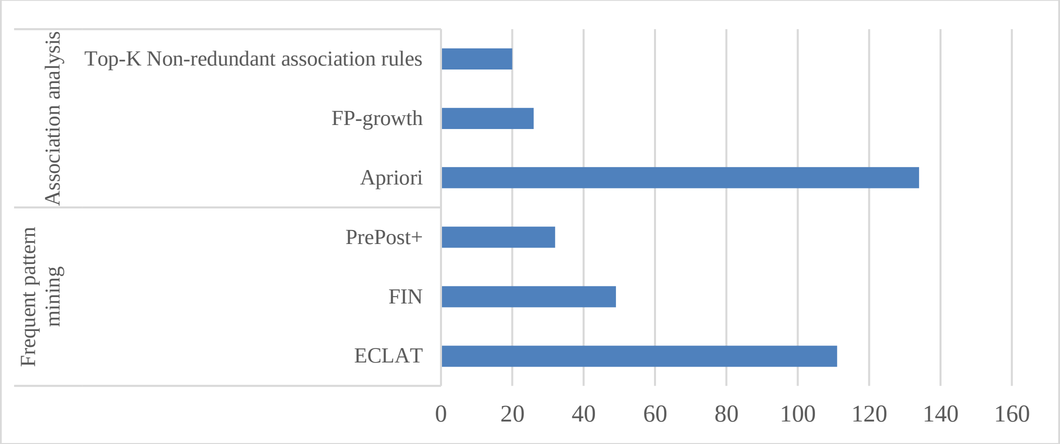


Figure 9 Runtime test of association analysis algorithms

The frequent pattern growth algorithm searches the entire database twice. In the first search, a list of items is created in descending order of appearance frequency by removing the items that appear below the minimum occurrence frequency for each item set in each row. In the second database search process, a frequent pattern tree is created by using the item set of the entire database based on the list as the created items. Once the target item is determined, the frequent pattern tree generated based on the target item is searched backwards and the association is analyzed.

2.3. Regression analysis

This chapter discusses how regression analysis can be applied to data mining of information from an offshore structure.

2.3.1. Assistance of purchase process by forecasting material requirement

Although there are various methods for applying regression analysis to data mining of offshore structures, this study tries to go in the direction of helping procurement and purchasing process. First, we analyze the procurement process that arises from the construction of offshore structures. After examining how regression analysis can help in this process, we examine what regression analysis can be applied.

(1) Method of assistance of purchase process

Figure 10 shows the procurement process that the shipyard mainly takes in building offshore structures. First, the offshore structure is divided into several modules after FEED stage. A production schedule is set up for each module, and a list of materials used for each module is created. A material requirements plan can be established by combining the production schedule and list of materials, and the material purchase plan can be established by applying the separately confirmed lead time to the material requirement plan.

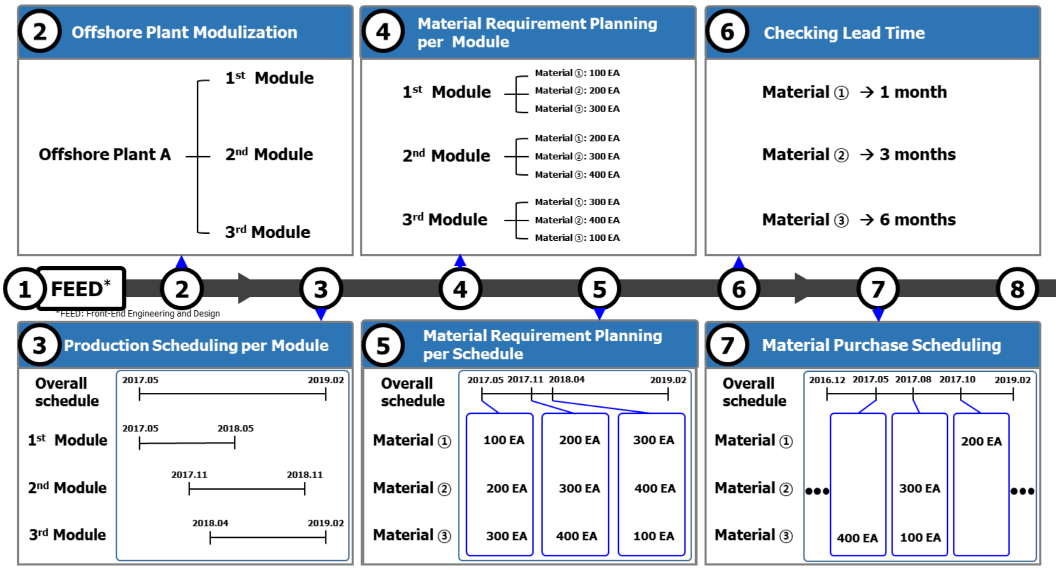
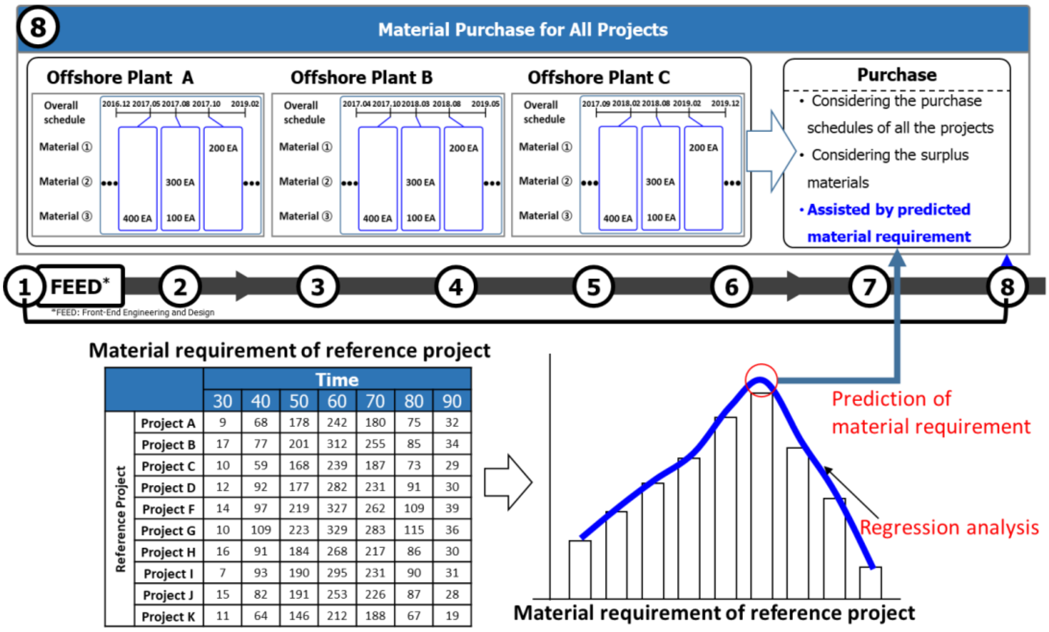


Figure 10 Procurement plan process of offshore structure

In the shipyard, the process described above should be carried out simultaneously for several projects as shown in Figure 11. Therefore, it may be difficult to formulate a plan for the same material that is required for several offshore structures at the same time. In this case, by using the data of the offshore structure, it is possible to predict the material requirements of the project to be carried forward by arranging the materials already consumed by the process schedule and creating the regression model. We think it can help procurement activities.


8
Material Purchase for All Projects
Offshore Plant A
Offshore Plant B
Offshore Plant C
Overall schedule
Material ①
Material ②
Material ③
2016.12 2017.05 2017.08 2017.10 2019.02
2017.04 2017.10 2018.03 2018.08 2019.05
2017.09 2018.02 2018.08 2019.02 2019.12
400 EA
300 EA
100 EA
200 EA
Purchase
• Considering the purchase schedules of all the projects
• Considering the surplus materials
• Assisted by predicted material requirement
1
FEED*
2
3
4
5
6
7
8
*FEED: Front-End Engineering and Design
Prediction of material requirement
Regression analysis
Material requirement of reference project

Material requirement of reference project

		Time						
		30	40	50	60	70	80	90
Reference Project	Project A	9	68	178	242	180	75	32
	Project B	17	77	201	312	255	85	34
	Project C	10	59	168	239	187	73	29
	Project D	12	92	177	282	231	91	30
	Project F	14	97	219	327	262	109	39
	Project G	10	109	223	329	283	115	36
	Project H	16	91	184	268	217	86	30
	Project I	7	93	190	295	231	90	31
	Project J	15	82	191	253	226	87	28
	Project K	11	64	146	212	188	67	19

Figure 11 Necessity of prediction of material requirement

(2) Principal of regression analysis

A regression analysis is a set of statistical processes for estimating the relationship among variables. It is basically about between a dependent variable and one or more independent variables. The regression analysis is helpful for statistical prediction such as

a. Time variable data

b. Result of theoretical experiment

c. Modeling of cause and effect relationship

Equation 1 is basic function of a regression analysis and main purpose of regression analysis is to find a line which minimize sum of squares residual which is $\boldsymbol{RSS} = \boldsymbol{e}_1^2 +$

$\cdots + \boldsymbol{e_n^2}$

$$Y_i = \beta_0 + \beta_i X_i + \varepsilon_i$$ Equation 1

Figure 12 shows the basic concept of regression analysis.

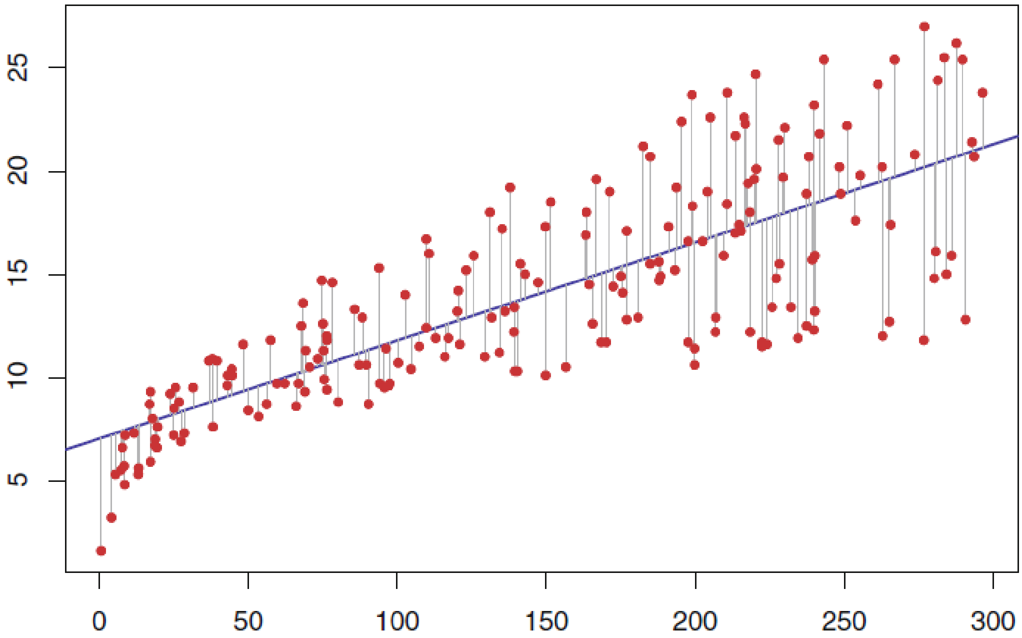


Figure 12 Concept of regression analysis

2.3.2. Regression model training

There are several ways to construct a regression model. In this chapter, we describe the regression model we tried in this study and explain how we constructed the regression model.

(1) Regression analysis in time series

The easiest way to predict material requirements through regression analysis is to analyze trends over time. Table 9 summarizes the material requirements when the process progresses for a certain period for each offshore structure. This table can be drawn by plotting the graph (Figure 13) with the horizontal axis as the time axis for the regression analysis. You can draw a line that minimizes the error value at each point, then it is possible to identify the trends of the offshore structure materials used at each time point. Advantage of predicting in time series is that we can predict material requirements even when precise requirements are not specified. But there is also disadvantage; various specifications and characteristics of each offshore structures cannot be reflected.

Table 9 Material consumption of offshore structures

		Time						
		30	**40**	**50**	**60**	**70**	**80**	**90**
Reference Project	**Project A**	9	68	178	242	180	75	32
	Project B	17	77	201	312	255	85	34
	Project C	10	59	168	239	187	73	29
	Project D	12	92	177	282	231	91	30
	Project F	14	97	219	327	262	109	39
	Project G	10	109	223	329	283	115	36
	Project H	16	91	184	268	217	86	30
	Project I	7	93	190	295	231	90	31
	Project J	15	82	191	253	226	87	28
	Project K	11	64	146	212	188	67	19

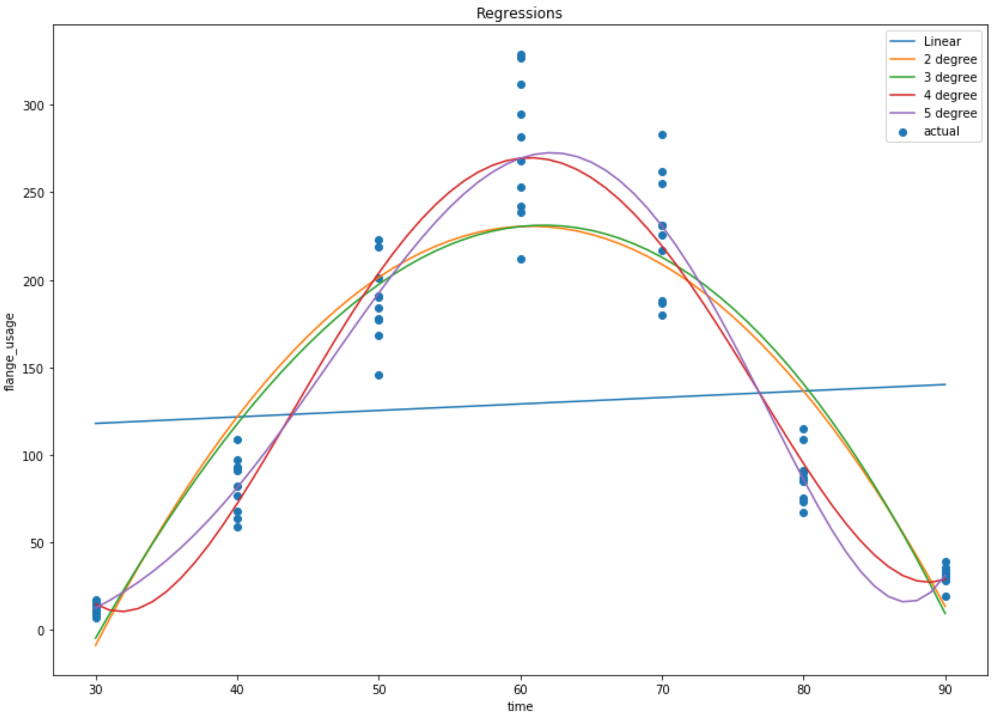


Figure 13 Material consumption trend of offshore structures

(2) Regression analysis for project characteristics

In order to overcome the shortcomings of the time-domain regression analysis, the independent variables of the regression analysis should be come from the characteristics of offshore structures. The regression model is composed only n-order polynomial. The advantage of regression analysis using characteristics of offshore structures as independent variables is that it is possible to predict various offshore structures as characteristics can be implemented. Disadvantage is that it is impossible to predict where requirement data for the desired point in time is not exist. Specification and related figures for each offshore structure are used as independent variables. One of the variables is main dimension of an offshore structure such as length, depth and draft. Also, we used storage capacity and production capacity including oil, gas and water. Other variables are total light weight of the offshore structure and number of people on board. The depth of well is also used as our independent variable. Table 10 lists the various independent variables used in this study and their values.

(3) Regression model by neural network

A regression model can be constructed using an artificial neural network. For comparison with traditional regression models, we construct a regression model by constructing an artificial neural network and compare the results. The neural network consists of a fully connected layer. The activation function is Relu, the hidden layer is 2, and the node is 15 for each layer. The input and output values are constructed as regression models. The input values are defined as material requirements for offshore structure and period, and material requirements for process period as output values. Figure 14 is showing the structure of artificial neural network.

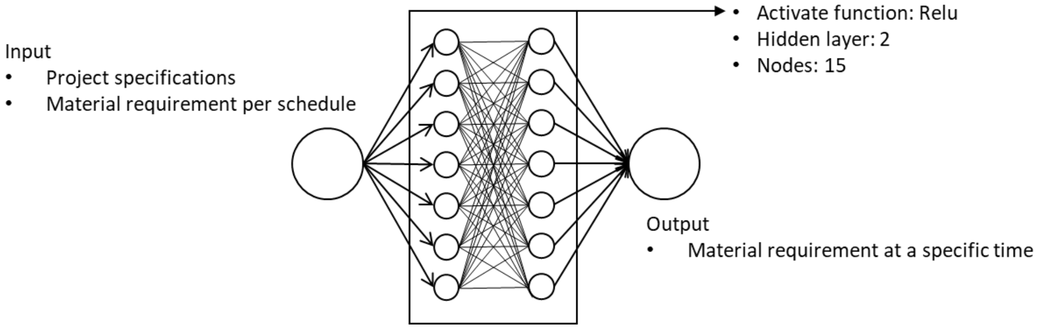

Input
• Project specifications
• Material requirement per schedule
• Activate function: Relu
• Hidden layer: 2
• Nodes: 15
Output
• Material requirement at a specific time

Figure 14 Structure of neural network

Table 10 Independent variables

Project	L (m)	B (m)	D (m)	T (m)	DWT (ton)	SC (MMBBL)	OP (MMCFD)	GP (MMCFD)	WP (MMBWPD)	CREW (person)	WD (mm)	TLWT (ton)
Project 1	214.7	38	23.7	18	88326	0.56	0.06	12	0.25	50	105	6200
Project 2	214.7	38.2	22.2	16	60000	0.42	0.057	53	0.057	77	84	2500
Project 3	232	41.5	23.5	16	92800	0.58	0.089	38	0.122	60	126	6000
Project 4	273	50	28	18.8	180000	1.4	0.003	35	0.174	84	366	11000
Project 5	276.4	45	26.6	19.7	105000	0.94	0.22	840	0.063	116	320	15000
Project 6	300	59.6	30.5	22.8	343000	2	0.27	280	0.18	140	1400	23500
Project 7	285	60	32.3	24.4	340660	2.2	0.25	400	0.525	100	1180	23000
Project 8	296	63	32.3	24	375600	2.2	0.21	340	0.15	100	1220	30000
Project 9	312.4	60	33.2	24.3	329000	2	0.24	280	0.265	190	1365	30000
Project 10	305.1	58	32	23.4	312500	2	0.225	170	0.1	70	1030	22000
Project 11	260	46	25.8	18.5	142000	0.9	0.1	80	0.1	80	390	8000
Project 12	319	58	31	23.4	360000	1.77	0.24	400	0.45	120	1310	24000
Project 13	320	58.4	32	24	337859	2.2	0.25	450	0.12	100	1462	35000
Project 14	310	61	30.5	23.5	321000	2	0.225	530	0.42	240	1325	37000
Project 15	320	61	32	24.7	353200	2	0.18	176.6	0.1	180	750	27700
Project 16	250.2	34	19.1	12.8	43276	0.28	0.14	100	0.12	70	450	4500
Project 17	253	42	23.2	15	103000	0.6	0.07	110	0.022	55	85	5000
Project 18	334.9	43.7	27.7	21.4	228033	1.5	0.1	52	0.02	76	383	3500
Project 19	245.4	39.6	20.6	14.7	94238	0.65	0.035	100	0.018	85	75	1900
Project 20	362	60	28.3	23	356400	1.3	0.081	75	0.05	100	700	4500
Project 21	271.8	46	26.6	18	150000	0.94	0.13	150	0.18	80	120	12000
Project 22	271	44	22.4	17	138900	1.04	0.08	85	0.032	100	70	5500
Project 23	337	54.5	27	21	273191	2	0.15	162	0.2	100	785	14000
Project 24	328.6	54.5	27	21	273622	1	0.15	210	0.251	194	1035	14000
Project 25	217.2	38	23	17	85943	0.45	0.1	75	0.3	90	113	6000
Project 26	325	61	32.5	25.6	320000	1.9	0.22	150	0.382	240	800	32000
Project 27	346.3	57.3	28.5	22.9	322911	2	0.14	35	0.325	46	1200	14000
Project 28	295	50.6	29	19.9	128000	0.95	0.085	671	0.02	126	350	16000
Project 29	271.7	46	26.6	18.2	148192	0.95	0.08	53	0.06	89	134	12000
Project 30	242.3	42	21.1	14.9	105000	0.66	0.06	85	0.065	96	350	4500
Project A	285	60	32.3	24.4	340,660	2.2	0.25	400	0.525	100	1,010	23,000
Project B	318.8	56	29.5	19.8	255,271	1.6	0.18	71	0.232	110	1,260	14,500
Project C	305	61	32	24	350,000	2	0.22	250	0.319	240	1,200	37,478

Artificial neural network is used by learning by using input value and output value. Figure 15 shows the loss value of learning the artificial neural network for the flange.

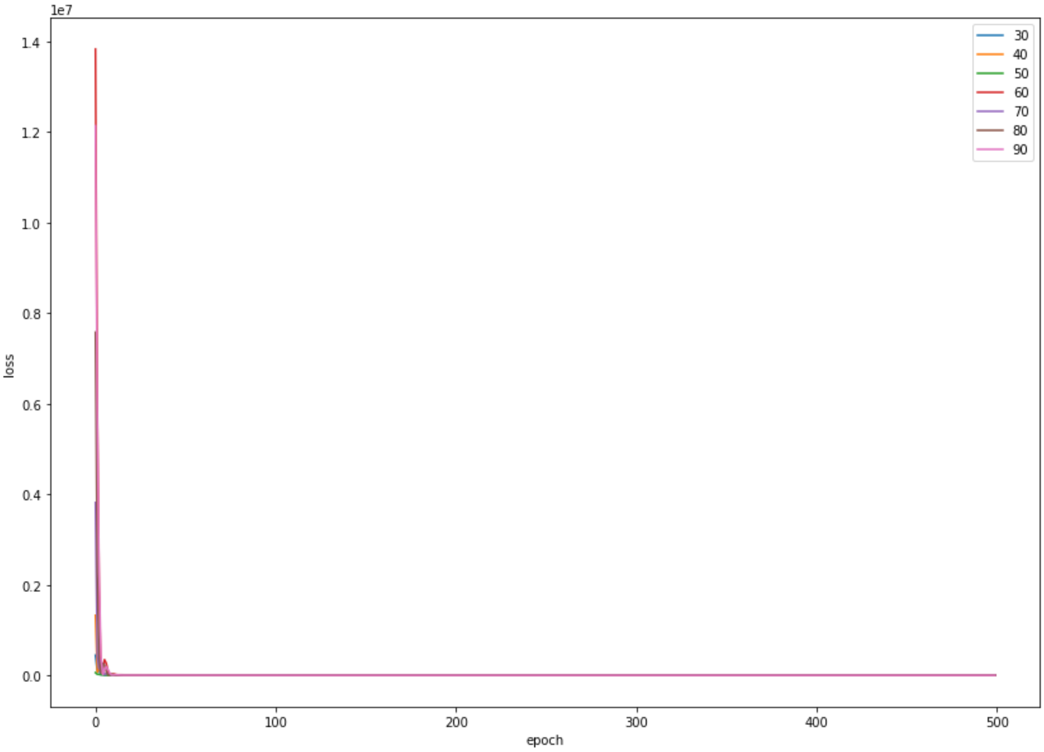


Figure 15 Loss during training of neural network

Figure 16 shows the prediction of the result using the artificial neural network learned on the flange of schedule 50

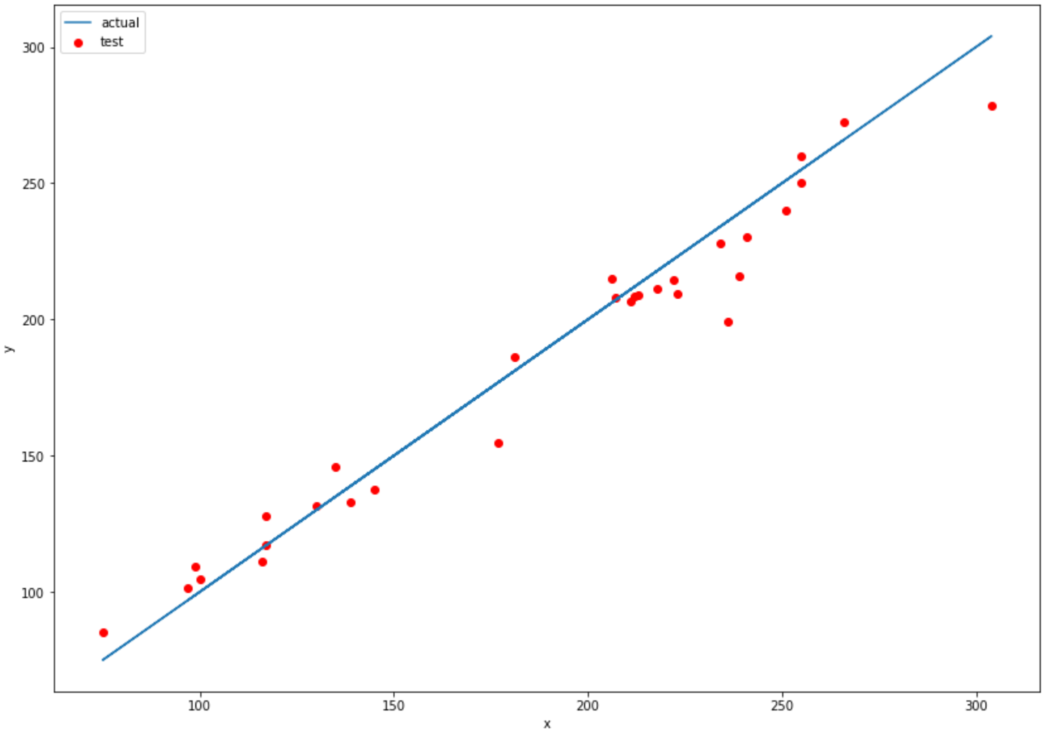


Figure 16 Prediction result of neural network

(4) Comparison of different regression models

Regression analysis varies according to the degree of each independent variable. Equation 2 shows a simple polynomial regression equation. Equation 3 shows a quadratic regression, and Equation 4 shows a cubic regression. In addition, we can make regression equation using log or exponential function.

$$\begin{aligned}&MaterialRequirement = \\ &x_0 \times L + x_1 \times B + x_2 \times D + x_3 \times T + x_4 \times DWT + \\ &x_5 \times SC + x_6 \times OP + x_7 \times GP + x_8 \times WP + \\ &x_9 \times CREW + x_{10} \times WD + x_{11} \times TLWT + x_{12}\end{aligned}$$

Equation 2

$$\begin{aligned}&MaterialRequirement = \\ &x_0 \times L + x_1 \times L^2 + x_2 \times B + x_3 \times B^2 + x_4 \times D + x_5 \times D^2 + \\ &x_6 \times T + x_7 \times T^2 + x_8 \times DWT + x_9 \times DWT^2 + x_{10} \times SC + x_{11} \times SC^2 + \\ &x_{12} \times OP + x_{13} \times OP^2 + x_{14} \times GP + x_{15} \times GP^2 + x_{16} \times WP + x_{17} \times WP^2 + \\ &x_{18} \times CREW + x_{19} \times CREW^2 + x_{20} \times WD + x_{21} \times WD^2 + \\ &x_{22} \times TLWT + x_{23} \times TLWT^2 + x_{24}\end{aligned}$$

Equation 3

$$\begin{aligned}&MaterialRequirement = \\ &x_0 \times L + x_1 \times L^2 + x_2 \times L^3 + x_3 \times B + x_4 \times B^2 + x_5 \times B^3 + \\ &x_6 \times D + x_7 \times D^2 + x_8 \times D^3 + x_9 \times T + x_{10} \times T^2 + x_{11} \times T^3 + \\ &x_{12} \times DWT + x_{13} \times DWT^2 + x_{14} \times DWT^3 + \\ &x_{15} \times SC + x_{16} \times SC^2 + x_{17} \times SC^3 + x_{18} \times OP + x_{19} \times OP^2 + x_{20} \times OP^3 + \\ &x_{21} \times GP + x_{22} \times GP^2 + x_{23} \times GP^3 + x_{24} \times WP + x_{25} \times WP^2 + x_{26} \times WP^3 + \\ &x_{27} \times CREW + x_{28} \times CREW^2 + x_{29} \times CREW^3 + \\ &x_{30} \times WD + x_{31} \times WD^2 + x_{32} \times WD^3 + x_{33} \times TLWT + x_{34} \times TLWT^2 + x_{35} \times TLWT^3 + x_{36}\end{aligned}$$

Equation 4

However, for the regression equation to have a correct predictive value, correlation analysis should be performed for each variable and it should be confirmed whether there is any correlation. Figure 17 shows the independent variables used in this study plotted against each other, and the Table 11 summarizes the correlation values. The correlation analysis used here is Pearson correlation coefficient.

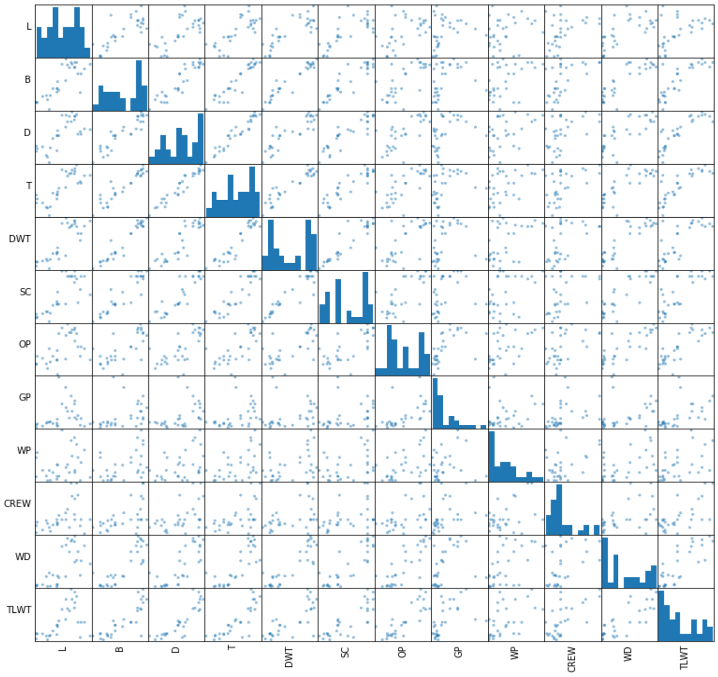


Figure 17 Correlation analysis

Table 11 Pearson correlation coefficient

	L	B	D	T	DWT	SC	OP	GP	WP	CREW	WD	TLWT
L	1.00	0.80	0.71	0.79	0.83	0.76	0.50	0.25	0.22	0.44	0.69	0.52
B		1.00	0.92	0.94	0.96	0.92	0.71	0.38	0.46	0.58	0.86	0.83
D			1.00	0.96	0.88	0.90	0.74	0.45	0.45	0.52	0.79	0.86
T				1.00	0.94	0.91	0.74	0.39	0.50	0.55	0.82	0.82
DWT					1.00	0.93	0.71	0.25	0.48	0.49	0.89	0.77
SC						1.00	0.74	0.33	0.46	0.43	0.86	0.82
OP							1.00	0.59	0.53	0.51	0.83	0.84
GP								1.00	0.18	0.41	0.40	0.56
WP									1.00	0.41	0.56	0.53
CREW										1.00	0.48	0.68
WD											1.00	0.81
TLWT												1.00

As a result of the correlation analysis, there were cases where each independent variable had a linear relationship, but it was not certain whether there was any other nonlinear relationship. Therefore, the second and third polynomials were constructed, and their values were compared to confirm accuracy. Table 12 is a regression analysis of the flanges of the project A and summarizes the errors of the predicted values. As can be seen in the table, the linear regression shows the most accurate value.

Table 12 Prediction error of regression models for flange

Schedule	Error ratio (%)				
	Actual	Linear	2nd	3rd	DL
30	14	6.07	59.86	70.09	55.97
40	98	23.63	72.99	78.00	32.84
50	221	24.73	63.19	23.00	3.64
60	339	7.92	0.43	27.08	2.67
70	264	1.38	80.45	95.58	15.54
80	110	2.34	16.58	63.57	1.24
90	39	27.28	118.70	326.25	7.02
Average of error ratio		13.34	58.89	97.65	16.99

Regarding the pipe materials of the project A, regression analysis was performed to check the error value of the result. Table 13 summarizes the results. Again, linear regression showed the most accurate predictions.

Table 13 Prediction error of regression models for pipe

Schedule	Error ratio (%)				
	Actual	Linear	2nd	3rd	DL
30	67	2.09	4.99	0.18	7.29
40	537	8.61	19.06	51.21	6.76
50	1545	2.24	50.88	51.00	6.17
60	1947	14.48	29.67	38.61	3.12
70	1545	1.23	7.41	11.77	9.77
80	739	8.52	43.76	80.80	5.93
90	336	4.29	39.85	157.83	20.60
Average of error ratio		5.92	27.94	55.92	8.52

As shown in Figure 18, the prediction of the pipeline of the project A is plotted as a graph and it can be confirmed that the predicted value of the linear regression is accurate for almost all cases. However, the prediction using the artificial neural network was the most accurate at the 60% processing time, which is the largest requirement.

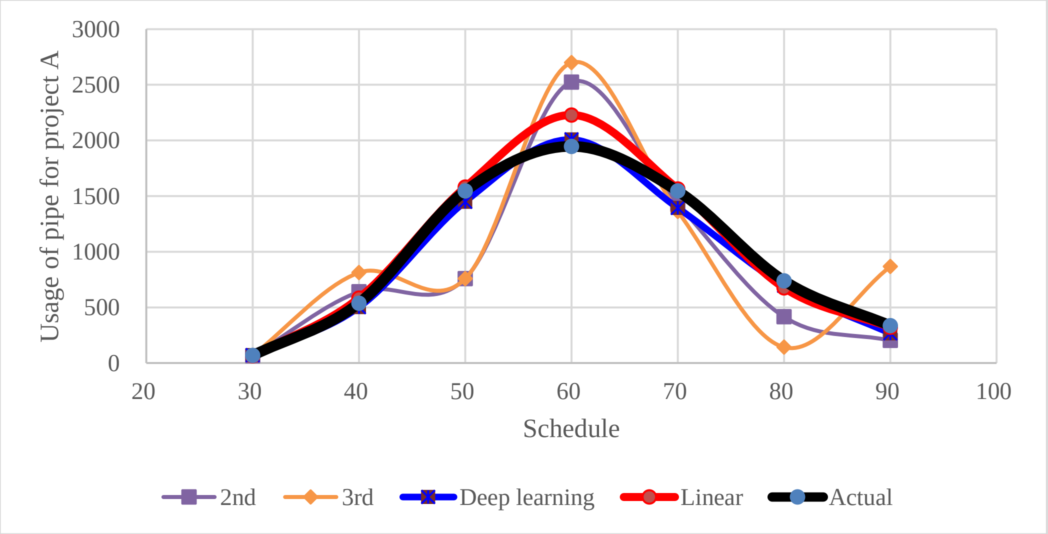


Figure 18 Prediction error for pipe

2.4. Estabilishment of big data framework

This chapter introduces the big data framework for applying the various data mining analysis methods described above.

2.4.1. Big data framework

The big data framework used in this study is based on the Hadoop ecosystem. Figure 19 shows the framework used in this study. The big data framework is based on Hadoop's HDFS (Hadoop Distributed File System), which supports the ability to distribute large files, a hive that supports functions such as relational databases, sparks that can implement various machine learning algorithms through distributed processing computing and zeppelin that can be configured as an analytical notebook environment. Here, the notebook environment refers to a work environment in which a screen capable of inputting anything such as a word processor on the Webpage, a code is written and executed, and a result is confirmed by repeating the result check and the code modification.

In addition, we use kylin, which to create data cubes based on HBase, which is free of additional input and possible to store data without index and store large amount of data. And, kylin is one of the OLAP tools supported by big data environment.

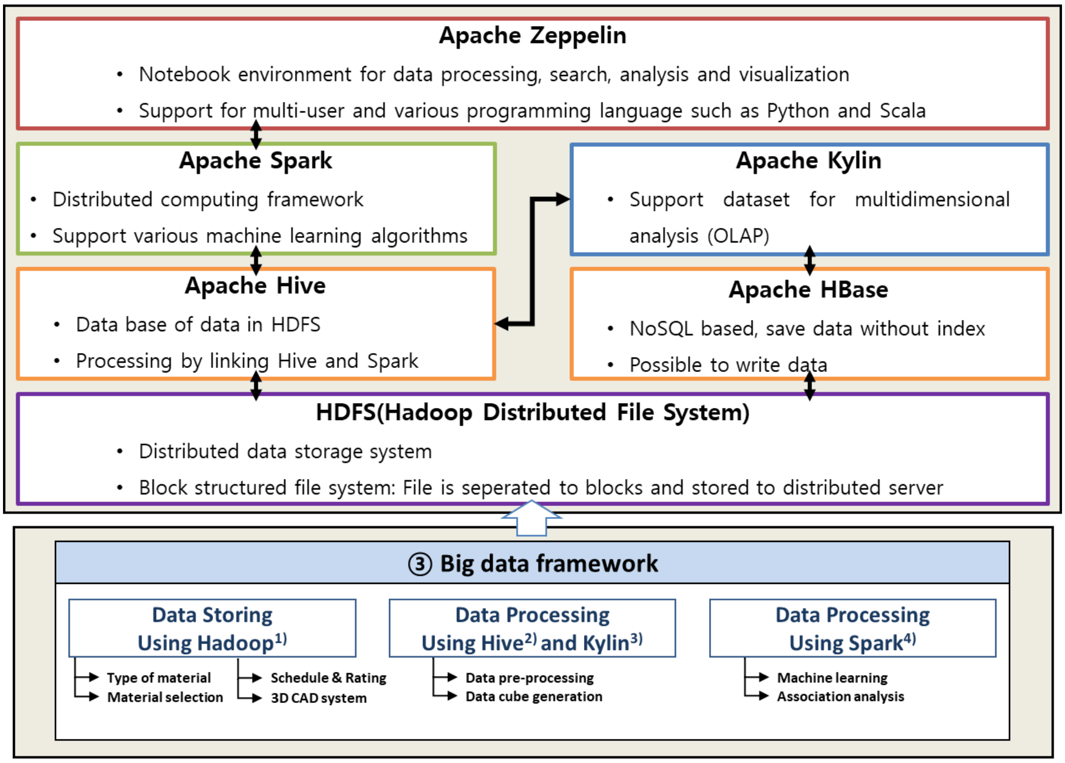

Apache Zeppelin
• Notebook environment for data processing, search, analysis and visualization
• Support for multi-user and various programming language such as Python and Scala
Apache Spark
• Distributed computing framework
• Support various machine learning algorithms
Apache Kylin
• Support dataset for multidimensional analysis (OLAP)
Apache Hive
• Data base of data in HDFS
• Processing by linking Hive and Spark
Apache HBase
• NoSQL based, save data without index
• Possible to write data
HDFS(Hadoop Distributed File System)
• Distributed data storage system
• Block structured file system: File is seperated to blocks and stored to distributed server
③ Big data framework
Data Storing Using Hadoop1)
Type of material
Material selection
Schedule & Rating
3D CAD system
Data Processing Using Hive2) and Kylin3)
Data pre-processing
Data cube generation
Data Processing Using Spark4)
Machine learning
Association analysis

Figure 19 Big data framework

2.4.2. Data processing

The piping material list includes the type, material, size and thickness of each material, and various reference materials required to complete the design. To do this, data cubes (multidimensional data) must be created from the file source to speed up data retrieval for the required content and apply search results to machine learning algorithms. The original data stored on the HDFS is transmitted as a hive to create a data table. The data table can be quickly retrieved for a desired item, and a data cube can be generated by extracting only necessary items by selecting a column to be included in the analysis.

Spark supports data mining through a variety of machine learning algorithms based on distributed processing computing capabilities. Python, and Scala to support data-mining techniques in a user-friendly environment. Associativity analysis and regression analysis can also be implemented on sparks, and the results can be viewed in real-time, And the results of the regression model can be obtained. A typical data pre-processing is summarized in Figure 20.

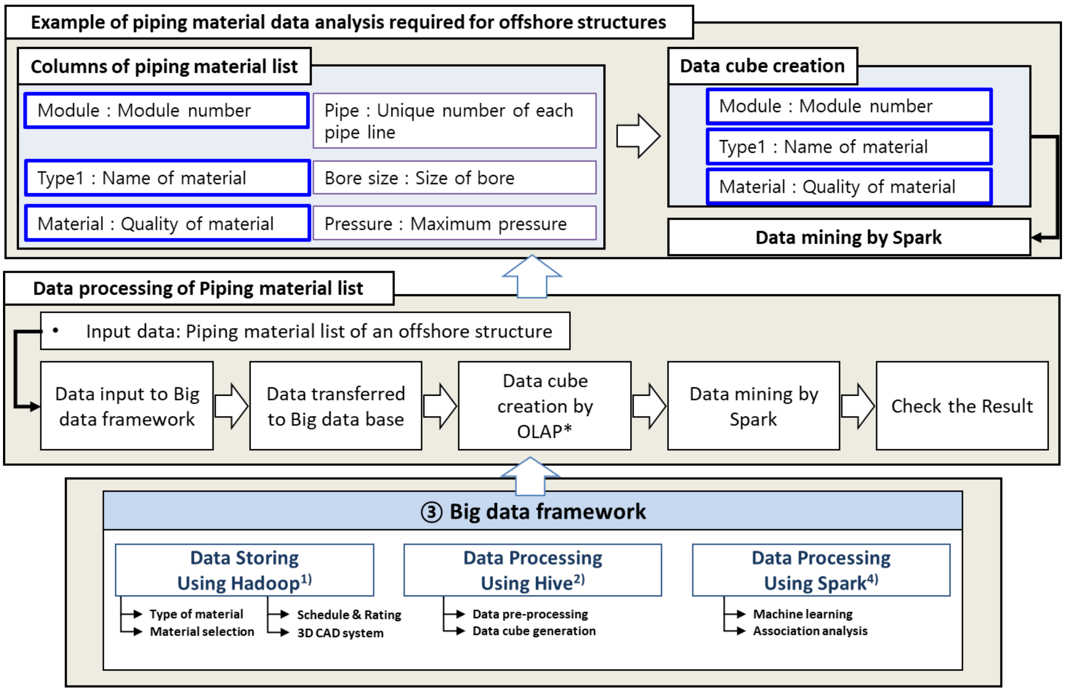

Example of piping material data analysis required for offshore structures
Columns of piping material list
Module : Module number
Pipe : Unique number of each pipe line
Type1 : Name of material
Bore size : Size of bore
Material : Quality of material
Pressure : Maximum pressure
Data cube creation
Module : Module number
Type1 : Name of material
Material : Quality of material
Data mining by Spark
Data processing of Piping material list
• Input data: Piping material list of an offshore structure
Data input to Big data framework
Data transferred to Big data base
Data cube creation by OLAP*
Data mining by Spark
Check the Result
③ Big data framework
Data Storing Using Hadoop1)
Type of material
Material selection
Schedule & Rating
3D CAD system
Data Processing Using Hive2)
Data pre-processing
Data cube generation
Data Processing Using Spark4)
Machine learning
Association analysis

Figure 20 Data processing

3. Application to big data framework

In this chapter, we will explain the case of applying the data mining method described in Chapter 2 as an application of Big Data Framework.

3.1. Recommendation of associated materials

Using the association analysis of piping materials through association analysis, it is possible to establish the basis of a system that recommends related materials.

3.1.1. Method of association analysis of piping material

(1) Overview of association analysis of piping material

In the overall process (Figure 21), first, a branch is regarded as a shopping cart, and the association analysis is performed by using the list of the materials required for the information of the offshore structure, the specification and type of the material, and the pipeline or the branch. The analysis results can be verified against actual cases.

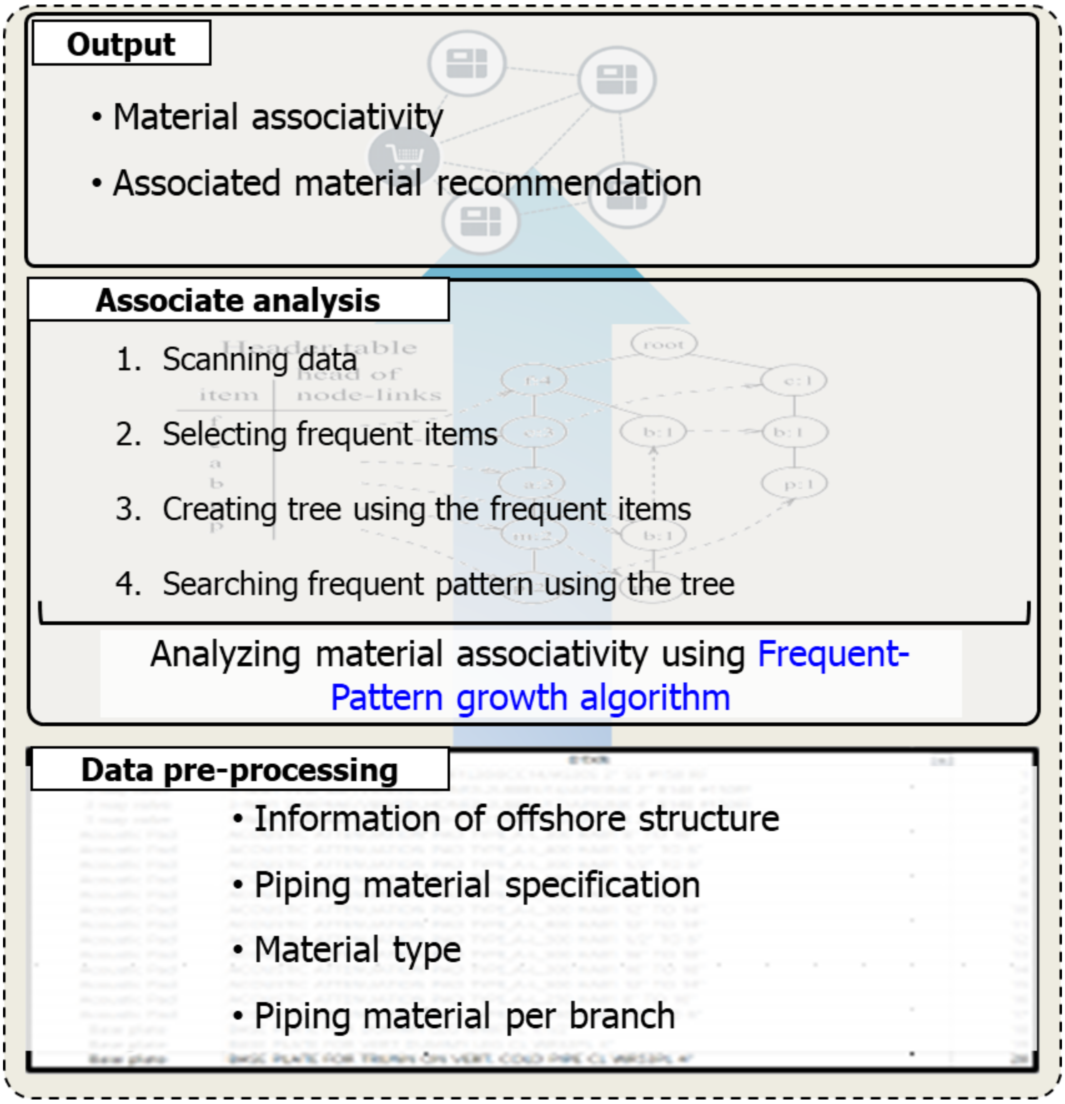
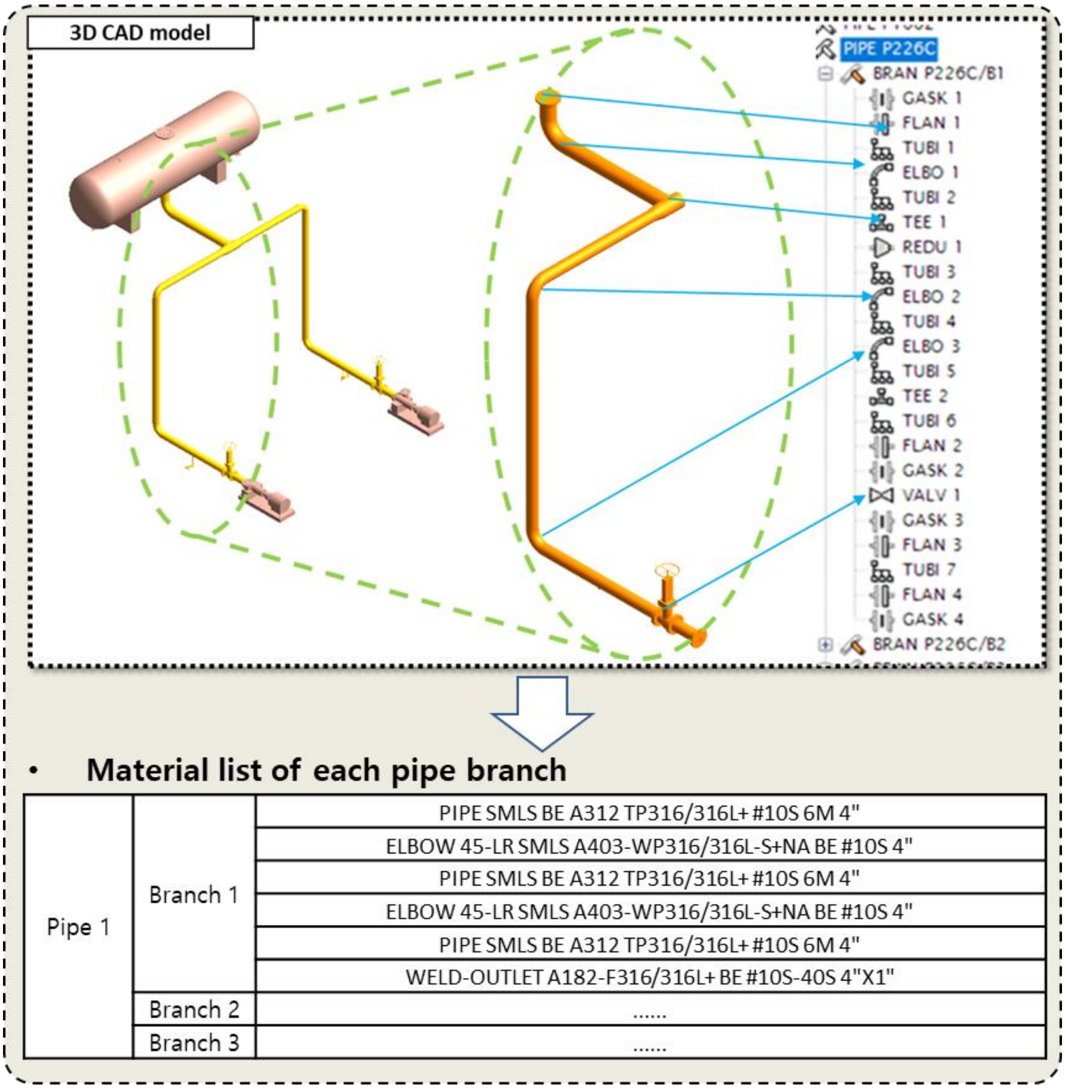

Output
• Material associativity
• Associated material recommendation
Associate analysis
1. Scanning data
2. Selecting frequent items
3. Creating tree using the frequent items
4. Searching frequent pattern using the tree
Analyzing material associativity using Frequent-Pattern growth algorithm
Data pre-processing
• Information of offshore structure
• Piping material specification
• Material type
• Piping material per branch

Figure 21 Overview of association analysis

(2) Input data of association analysis

The piping material consists of hierarchy in one branch on a three-dimensional design tool, and the designer can extract it into a list. Figure 22 shows this.



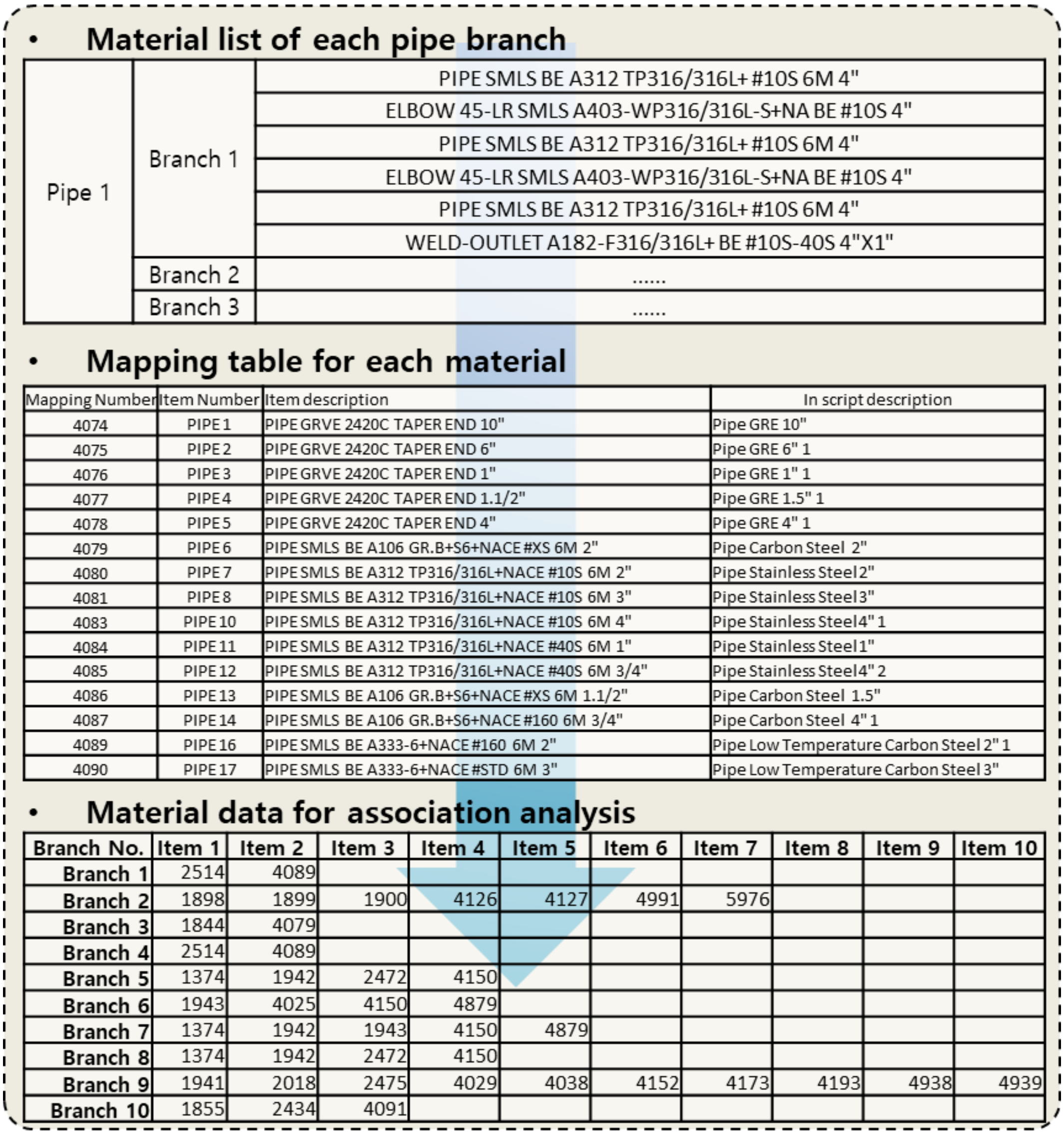
- **Material list of each pipe branch**

Pipe 1	Branch 1	PIPE SMLS BE A312 TP316/316L+ #10S 6M 4"
		ELBOW 45-LR SMLS A403-WP316/316L-S+NA BE #10S 4"
		PIPE SMLS BE A312 TP316/316L+ #10S 6M 4"
		ELBOW 45-LR SMLS A403-WP316/316L-S+NA BE #10S 4"
		PIPE SMLS BE A312 TP316/316L+ #10S 6M 4"
		WELD-OUTLET A182-F316/316L+ BE #10S-40S 4"X1"
	Branch 2	......
	Branch 3	......

Figure 22 Extracting material list from CAD tool

(3) Preprocessing of input data

The extracted list must be preprocessed since it cannot be directly input to the association analysis algorithm. Each material has a unique description for each material and a mapping number is created based on the description so that one unique number per material can be assigned. Since the shipyard usually uses material numbers for each material, it is safe to omit the mapping process when this study is applied at the shipyard. Figure 23 shows the process of generating and mapping unique numbers for each material.



- **Material list of each pipe branch**

Pipe 1	Branch 1	PIPE SMLS BE A312 TP316/316L+ #10S 6M 4"
		ELBOW 45-LR SMLS A403-WP316/316L-S+NA BE #10S 4"
		PIPE SMLS BE A312 TP316/316L+ #10S 6M 4"
		ELBOW 45-LR SMLS A403-WP316/316L-S+NA BE #10S 4"
		PIPE SMLS BE A312 TP316/316L+ #10S 6M 4"
		WELD-OUTLET A182-F316/316L+ BE #10S-40S 4"X1"
	Branch 2	......
	Branch 3	......

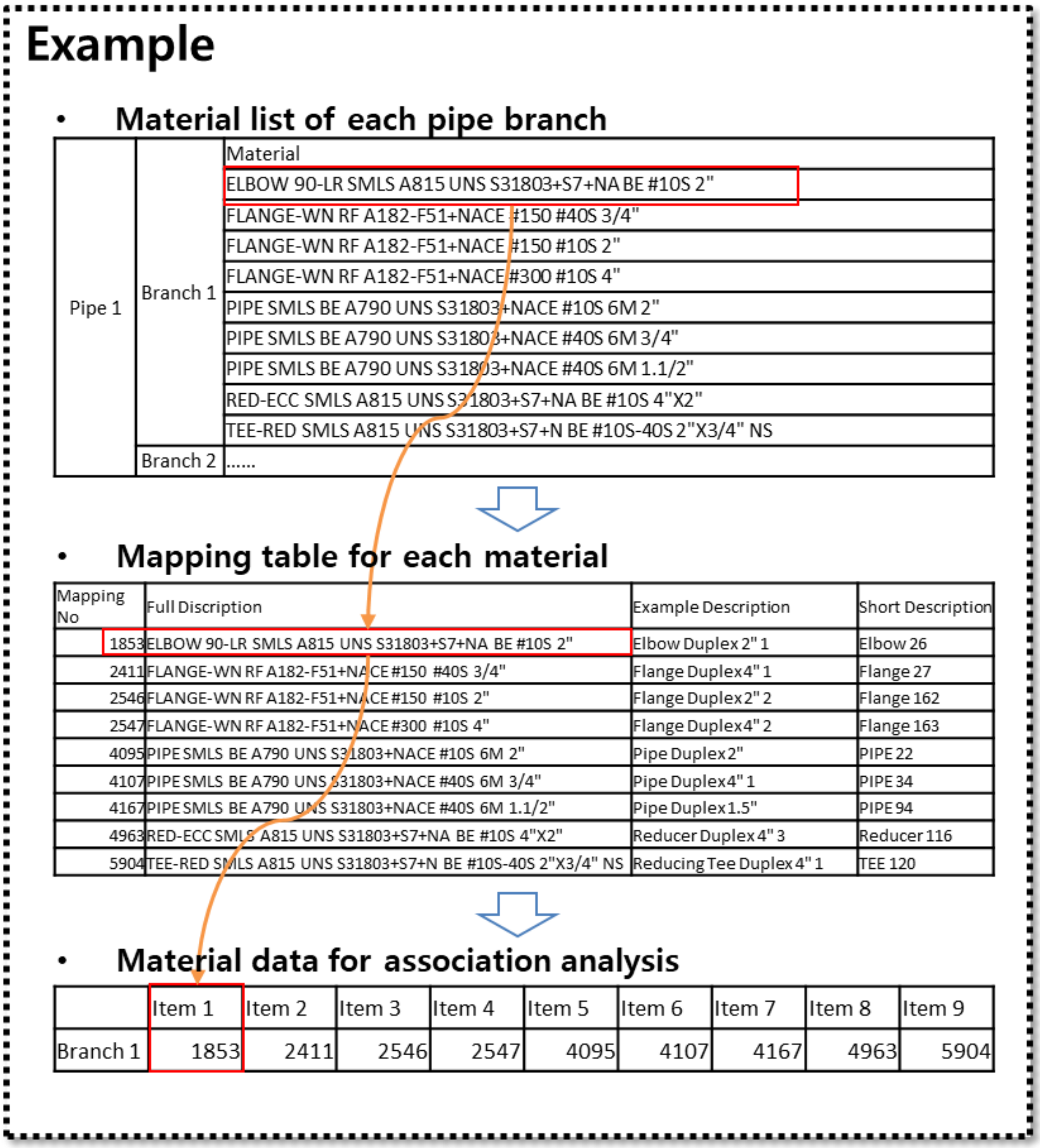
- **Mapping table for each material**

Mapping Number	Item Number	Item description	In script description
4074	PIPE 1	PIPE GRVE 2420C TAPER END 10"	Pipe GRE 10"
4075	PIPE 2	PIPE GRVE 2420C TAPER END 6"	Pipe GRE 6" 1
4076	PIPE 3	PIPE GRVE 2420C TAPER END 1"	Pipe GRE 1" 1
4077	PIPE 4	PIPE GRVE 2420C TAPER END 1.1/2"	Pipe GRE 1.5" 1
4078	PIPE 5	PIPE GRVE 2420C TAPER END 4"	Pipe GRE 4" 1
4079	PIPE 6	PIPE SMLS BE A106 GR.B+S6+NACE #XS 6M 2"	Pipe Carbon Steel 2"
4080	PIPE 7	PIPE SMLS BE A312 TP316/316L+NACE #10S 6M 2"	Pipe Stainless Steel 2"
4081	PIPE 8	PIPE SMLS BE A312 TP316/316L+NACE #10S 6M 3"	Pipe Stainless Steel 3"
4083	PIPE 10	PIPE SMLS BE A312 TP316/316L+NACE #10S 6M 4"	Pipe Stainless Steel 4" 1
4084	PIPE 11	PIPE SMLS BE A312 TP316/316L+NACE #40S 6M 1"	Pipe Stainless Steel 1"
4085	PIPE 12	PIPE SMLS BE A312 TP316/316L+NACE #40S 6M 3/4"	Pipe Stainless Steel 4" 2
4086	PIPE 13	PIPE SMLS BE A106 GR.B+S6+NACE #XS 6M 1.1/2"	Pipe Carbon Steel 1.5"
4087	PIPE 14	PIPE SMLS BE A106 GR.B+S6+NACE #160 6M 3/4"	Pipe Carbon Steel 4" 1
4089	PIPE 16	PIPE SMLS BE A333-6+NACE #160 6M 2"	Pipe Low Temperature Carbon Steel 2" 1
4090	PIPE 17	PIPE SMLS BE A333-6+NACE #STD 6M 3"	Pipe Low Temperature Carbon Steel 3"

- **Material data for association analysis**

Branch No.	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Branch 1	2514	4089								
Branch 2	1898	1899	1900	4126	4127	4991	5976			
Branch 3	1844	4079								
Branch 4	2514	4089								
Branch 5	1374	1942	2472	4150						
Branch 6	1943	4025	4150	4879						
Branch 7	1374	1942	1943	4150	4879					
Branch 8	1374	1942	2472	4150						
Branch 9	1941	2018	2475	4029	4038	4152	4173	4193	4938	4939
Branch 10	1855	2434	4091							

Figure 23 Mapping unique number to each material



Example

- Material list of each pipe branch

		Material
Pipe 1	Branch 1	ELBOW 90-LR SMLS A815 UNS S31803+S7+NA BE #10S 2"
		FLANGE-WN RF A182-F51+NACE #150 #40S 3/4"
		FLANGE-WN RF A182-F51+NACE #150 #10S 2"
		FLANGE-WN RF A182-F51+NACE #300 #10S 4"
		PIPE SMLS BE A790 UNS S31803+NACE #10S 6M 2"
		PIPE SMLS BE A790 UNS S31803+NACE #40S 6M 3/4"
		PIPE SMLS BE A790 UNS S31803+NACE #40S 6M 1.1/2"
		RED-ECC SMLS A815 UNS S31803+S7+NA BE #10S 4"X2"
		TEE-RED SMLS A815 UNS S31803+S7+N BE #10S-40S 2"X3/4" NS
	Branch 2	......

- Mapping table for each material

Mapping No	Full Discription	Example Description	Short Description
1853	ELBOW 90-LR SMLS A815 UNS S31803+S7+NA BE #10S 2"	Elbow Duplex 2" 1	Elbow 26
2411	FLANGE-WN RF A182-F51+NACE #150 #40S 3/4"	Flange Duplex4" 1	Flange 27
2546	FLANGE-WN RF A182-F51+NACE #150 #10S 2"	Flange Duplex2" 2	Flange 162
2547	FLANGE-WN RF A182-F51+NACE #300 #10S 4"	Flange Duplex4" 2	Flange 163
4095	PIPE SMLS BE A790 UNS S31803+NACE #10S 6M 2"	Pipe Duplex2"	PIPE 22
4107	PIPE SMLS BE A790 UNS S31803+NACE #40S 6M 3/4"	Pipe Duplex4" 1	PIPE 34
4167	PIPE SMLS BE A790 UNS S31803+NACE #40S 6M 1.1/2"	Pipe Duplex1.5"	PIPE 94
4963	RED-ECC SMLS A815 UNS S31803+S7+NA BE #10S 4"X2"	Reducer Duplex 4" 3	Reducer 116
5904	TEE-RED SMLS A815 UNS S31803+S7+N BE #10S-40S 2"X3/4" NS	Reducing Tee Duplex 4" 1	TEE 120

- Material data for association analysis

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
Branch 1	1853	2411	2546	2547	4095	4107	4167	4963	5904

Figure 24 Example of preprocessing

Figure 24 shows an example of how input values of association analysis are generated from a list of materials.

(4) Appling association algorithm

As soon as an input value that can be input to the association analysis is generated, the association analysis can be started. Because the pattern growth algorithm is used for the association analysis, the input values are scanned twice, and finally the material-specific associations are output as a result (Figure 25). The input values used in this study are summarized in Table 14. The results of the association analysis are as shown in Table 15, and examples are shown in Table 16.

Table 14 Number of inputs

No. of branch	No. of material types	No. of items
2591	449	35786

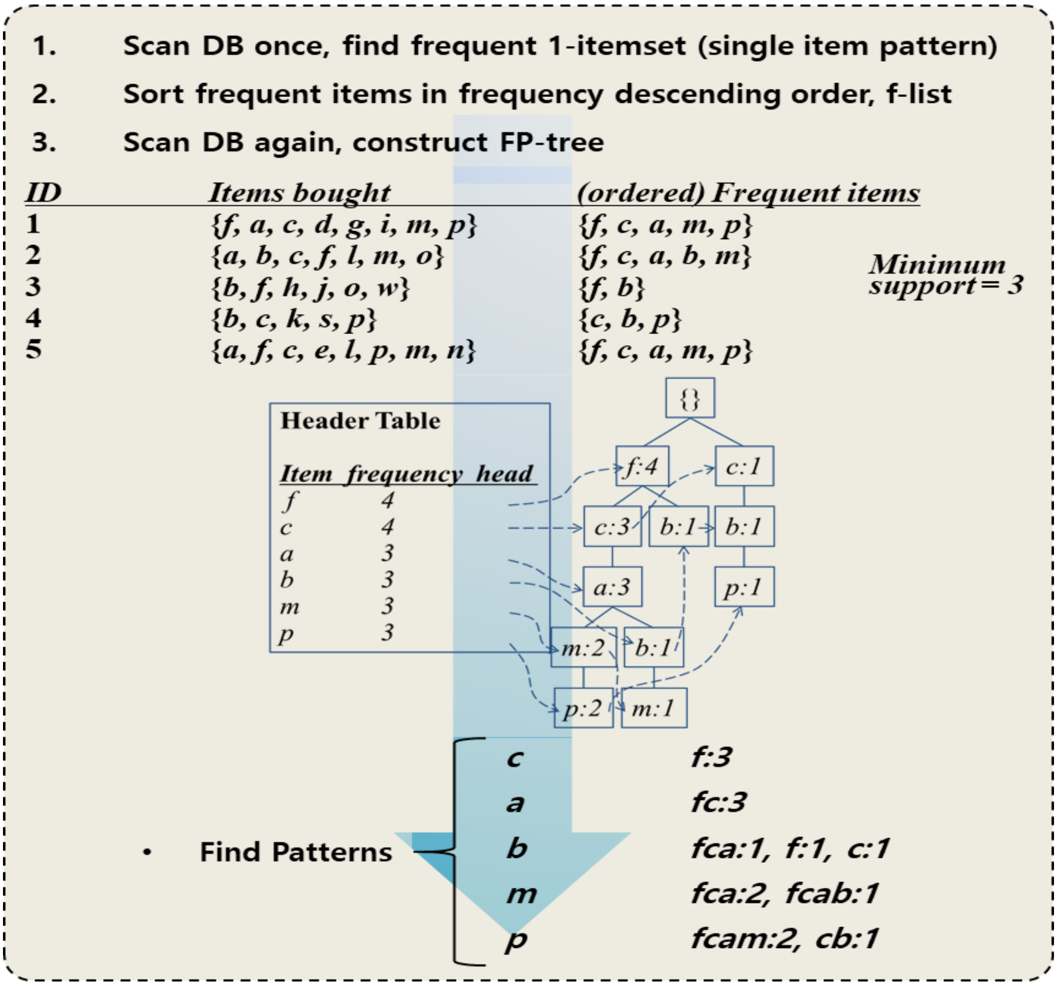
Table 15 Number of outputs

Association type	Count	Confidence mean
1:1	80	0.934363829
2:1	144	0.952661667
3:1	82	0.961845735
4:1	17	0.970306297
Total	323	0.951389919

Table 16 Example of outputs

Branch	Associations
Branch 1	[4085,1838,4080] => [5800], 1.0
Branch 2	[4085,1838,4080] => [2395], 1.0
Branch 3	[1374,4025,1942] => [4150], 1.0
Branch 4	[1830] => [4076], 1.0

Branch 5	[1855] => [4091], 0.9858156028368794
Branch 6	[4027,2474,1951] => [4148], 1.0
Branch 7	[2391,1838] => [4080], 0.9735849056603774
Branch 8	[1955] => [4078], 1.0
Branch 9	[1955] => [2388], 0.9733333333333334
Branch 10	[1374,1942,1943] => [4150], 1.0
Branch 11	[1374,1942,1943] => [2472], 0.8674698795180723
Branch 12	[1374,2472] => [4150], 1.0
Branch 13	[4869,4084] => [1906], 0.8688524590163934
Branch 14	[4869,4084] => [2391], 0.9344262295081968
Branch 15	[4118,1838] => [4080], 1.0



1. Scan DB once, find frequent 1-itemset (single item pattern)
2. Sort frequent items in frequency descending order, f-list
3. Scan DB again, construct FP-tree

ID	*Items bought*	*(ordered) Frequent items*
1	{*f, a, c, d, g, i, m, p*}	{*f, c, a, m, p*}
2	{*a, b, c, f, l, m, o*}	{*f, c, a, b, m*}
3	{*b, f, h, j, o, w*}	{*f, b*}
4	{*b, c, k, s, p*}	{*c, b, p*}
5	{*a, f, c, e, l, p, m, n*}	{*f, c, a, m, p*}

Minimum support = 3

Header Table

Item	*frequency*	*head*
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	

- Find Patterns

c	*f:3*
a	*fc:3*
b	*fca:1, f:1, c:1*
m	*fca:2, fcab:1*
p	*fcam:2, cb:1*

Figure 25 Process of FP growth algorithm

3.1.2. Result of analysis and validation

In order to verify the results of the association analysis, we compared the design model of the actual offshore structure and verified whether the material was used as a result of the correlation analysis.

(1) Relationship for one piping component

The 1: 1 association results are relatively simple. Most of the welded sets of materials appeared, and as a result, there was also a relatively low association. Table 17 shows the 1: 1 association results, and the actual examples can be found in Figure 26.

Table 17 Result of 1:1 association

	Used material	Recommended material	Confidence
1	Flange Carbon Steel 2inch 1	Pipe Carbon Steel 2inch	0.98
2	Tee Carbon Steel 2inch 1	Pipe Carbon Steel 2inch	0.89
3	Elbow Carbon Steel 2inch 2	Pipe Carbon Steel 2inch	0.96

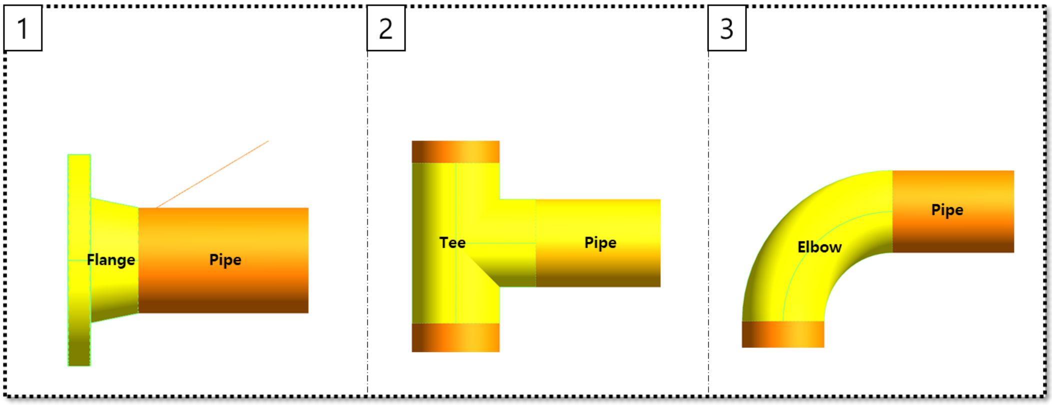


Figure 26 Figure of 1:1 association

(2) Relationship for two piping components

In the case of 1: 2 associativity, it is common to connect pipes of different sizes. In other words, the use of materials such as tee and the reducer are the result of the association analysis. Table 18 shows the results of the 1: 2 association analysis, and actual examples for each can be found in Figure 27 and Figure 28.

Table 18 Results of 1:2 associations

	Used material 1	Used material 2	Recommended material	Confidence
1	Pipe Carbon Steel 1.5inch	Reducer Carbon Steel 2inch	Pipe Carbon Steel 2inch	1
2	Elbow Carbon Steel 2inch 1	Elbow Carbon Steel 2inch 2	Pipe Carbon Steel 2inch	1
3	Reducing Tee Carbon Steel 4inch 1	Pipe Carbon Steel 4inch 1	Pipe Carbon Steel 2inch	1
4	Reducing Tee Carbon Steel 4inch 1	Elbow Carbon Steel 2inch 2	Pipe Carbon Steel 2inch	1

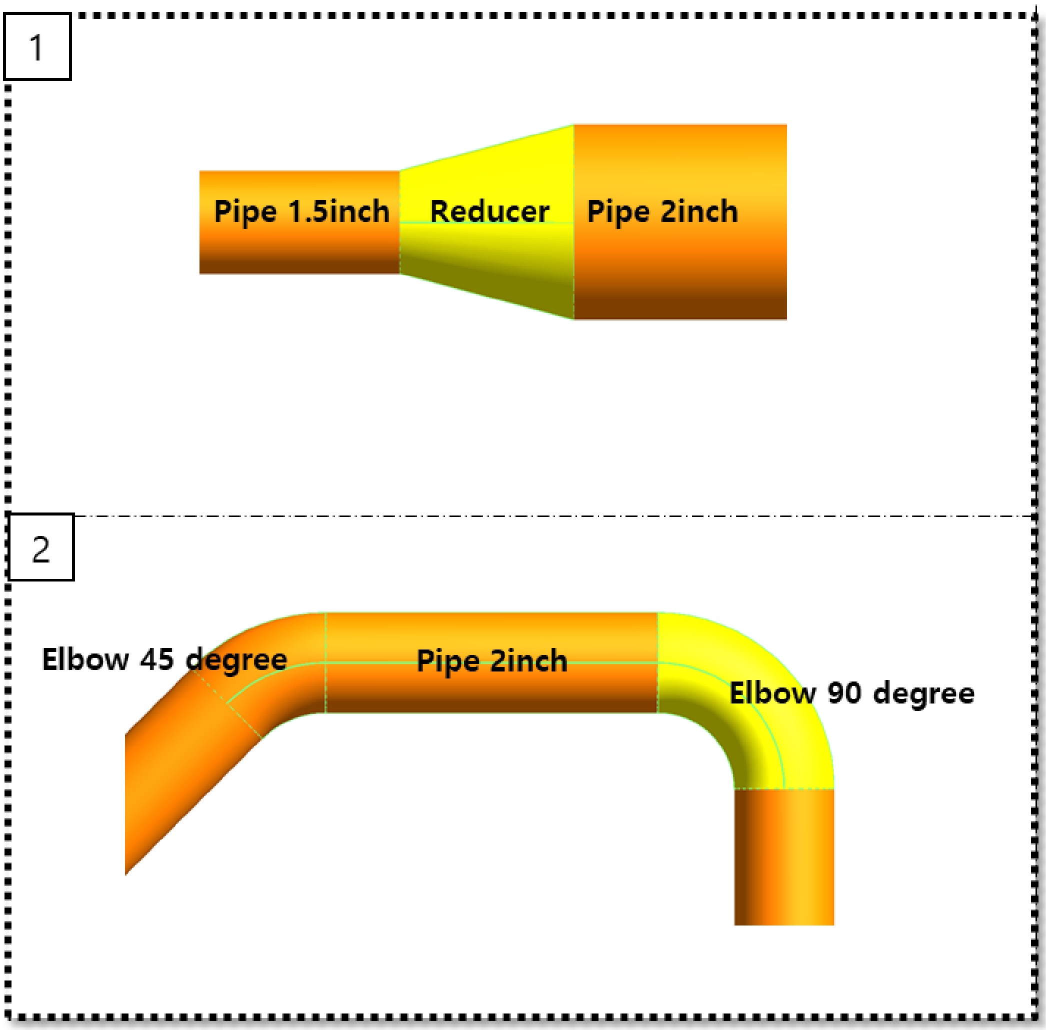


Figure 27 Figure of 1:2 associations 1

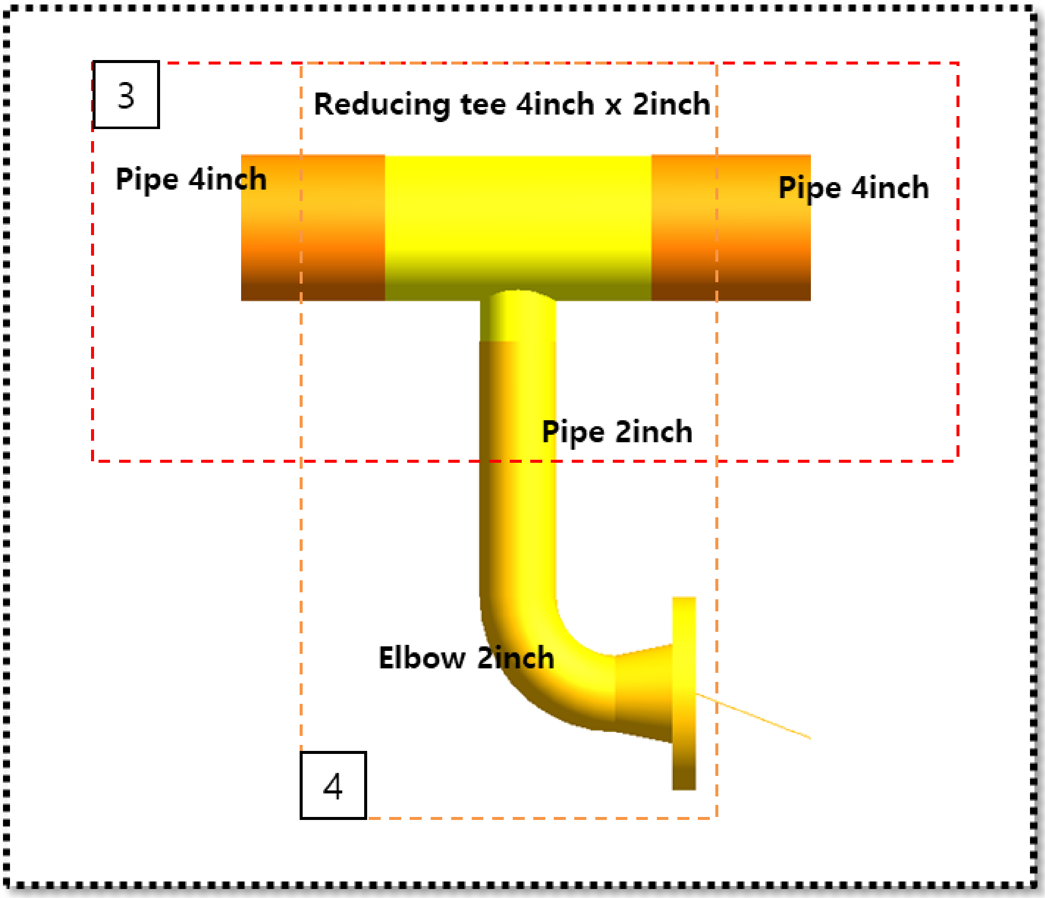


Figure 28 Figure of 1:2 associations 2

(3) Relationship for three and more piping components

For 1: 3 associations and further associations, we have shown a more complex shape of the piping design. As a result of the correlation analysis, it was confirmed that not only the size of piping, but also other materials were added. Table 19 shows the results of the 1: 3 association analysis, and Table 20 shows the results of the 1: 4 association analysis. Figure 29 shows the actual cases of 1: 3 association analysis results, and all the material sets shown in the same figure are 1: 4 associations.

Table 19 Result of 1:3 associations

	Used material 1	**Used material 2**	**Used material 3**	**Recommended material**	**Confidence**
1	Reducing Tee Carbon Steel 4inch 1	Flange Carbon Steel 4inch 1	Elbow Carbon Steel 2inch 2	Pipe Carbon Steel 2inch	1
2	Reducing Tee Carbon Steel 4inch 1	Pipe Carbon Steel 4inch 1	Flange Carbon Steel 4inch 1	Pipe Carbon Steel 2inch	1
3	Reducing Tee Carbon Steel 4inch 1	Pipe Carbon Steel 4inch 1	Elbow Carbon Steel 2inch 2	Pipe Carbon Steel 2inch	1

Table 20 Result of 1:4 association

	Used material 1	**Used material 2**	**Used material 3**	**Used material 4**	**Recommended material**	**Confidence**
*	Reducing Tee Carbon Steel 4inch 1	Pipe Carbon Steel 4inch 1	Flange Carbon Steel 4inch 1	Elbow Carbon Steel 2inch 2	Pipe Carbon Steel 2inch	1

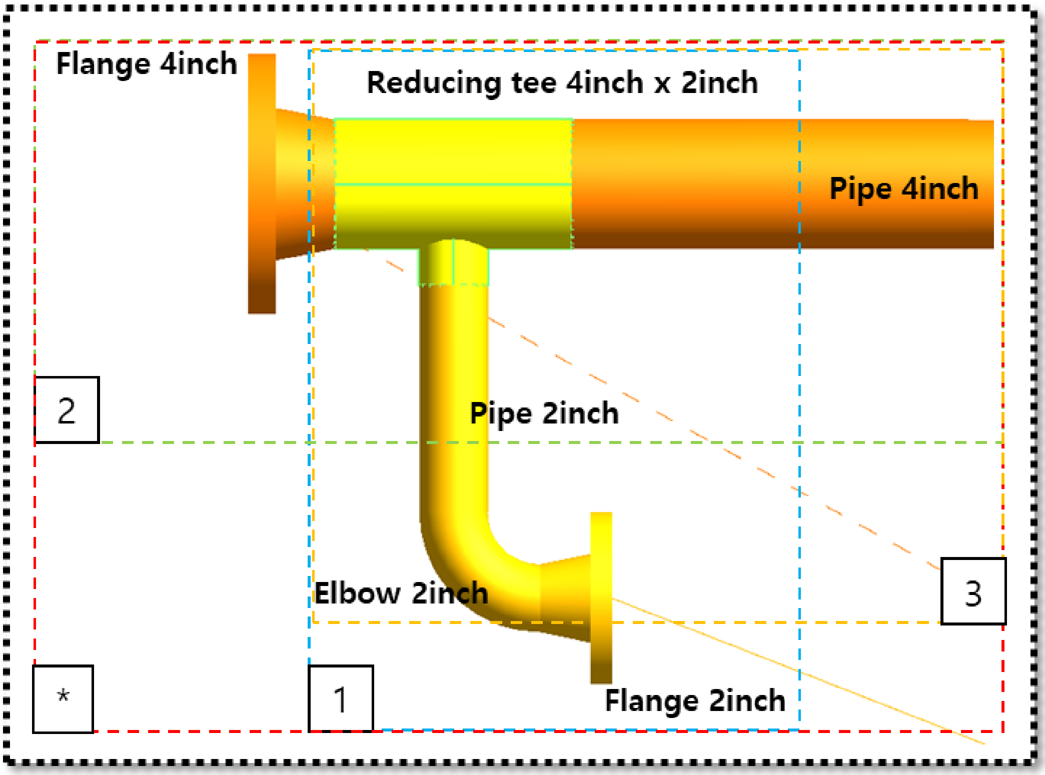


Figure 29 Example of 1:3 and more associations

(4) Different configuration using same piping components

In the present study, the values used as the input values of the correlation analysis did not include the position or orientation information of each material. Therefore, one of the disadvantages is the result of the same association, but there are cases where other design features are seen. As you can see in Table 21, all the combinations used the same material, but they all show different shapes as shown in Figure 30.

Table 21 Result of same association with different configuration

	Used material 1	**Used material 2**	**Recommended material**	**Confidence**
1	Tee Carbon Steel 4inch 1	Elbow Carbon Steel 4inch 1	Pipe Carbon Steel 4inch	1
2	Elbow Carbon Steel 4inch 1	Pipe Carbon Steel 4inch	Tee Carbon Steel 4inch 1	1
3	Pipe Carbon Steel 4inch	Tee Carbon Steel 4inch 1	Elbow Carbon Steel 4inch 1	1

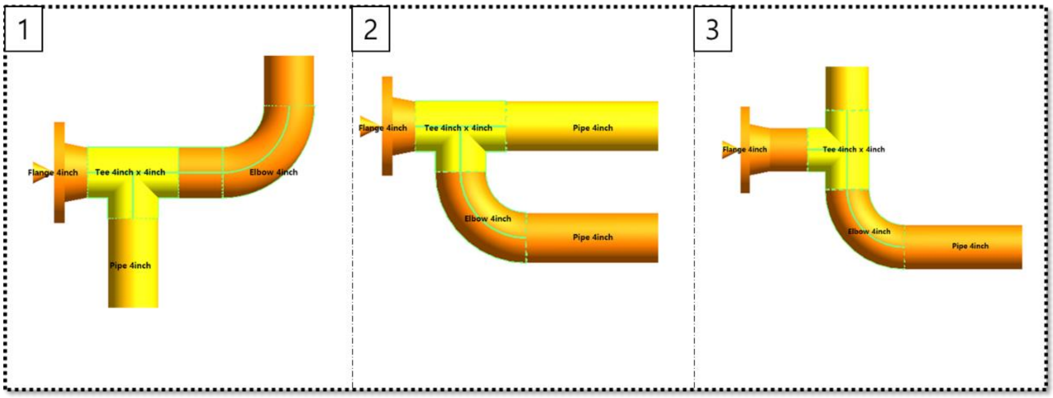


Figure 30 Example of same association with different configuration

3.2. Forecasting of material requirement

This chapter explains the prediction of material requirements as an application example of regression analysis.

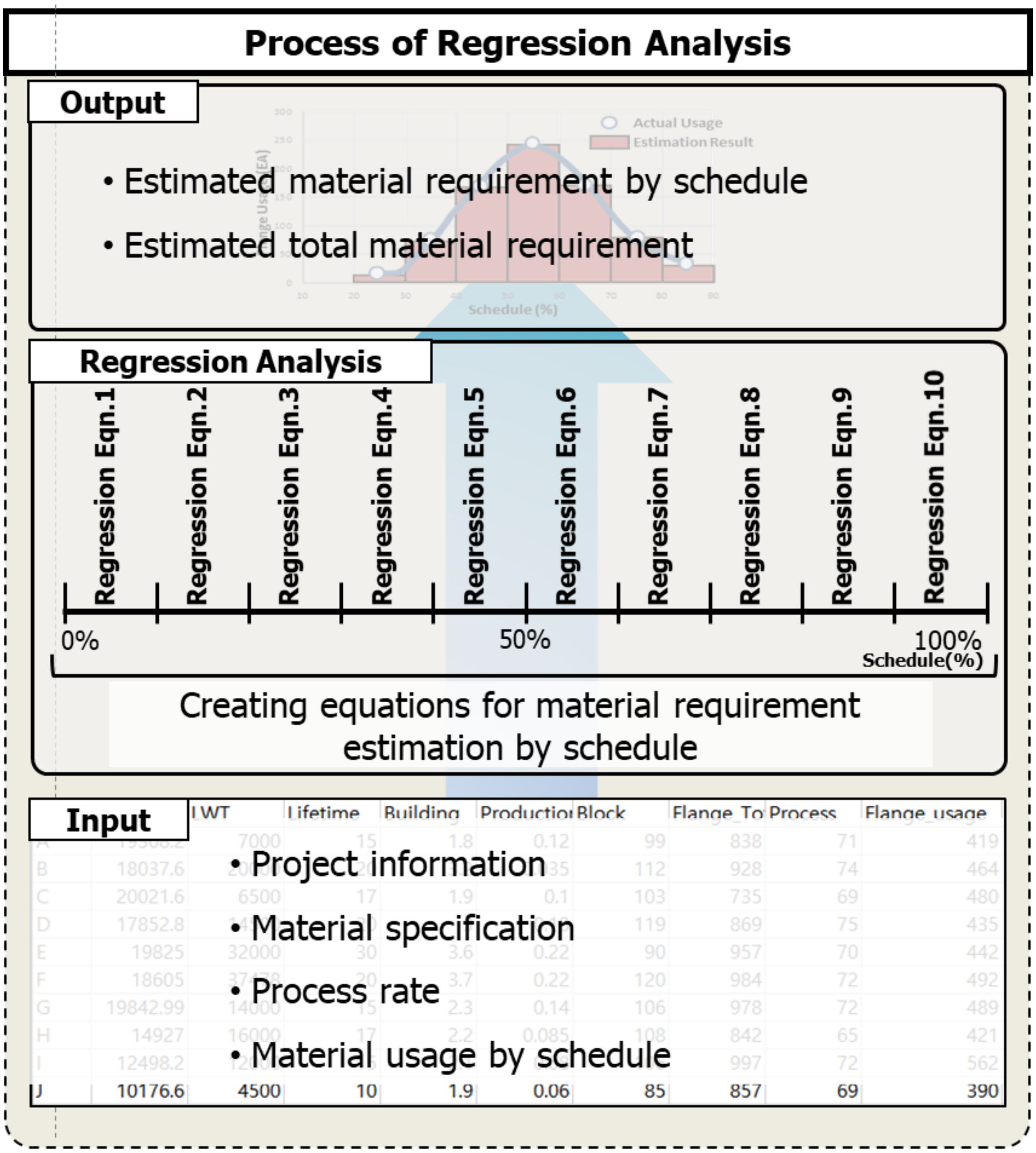
3.2.1. Method of training regression model

(1) Overview of regression analysis

The overall process of regression analysis is as follows. First, we check the material requirements by information of marine structure, material specification, process progress rate and process progress rate. Since each ocean structure has different air, the process is divided into 10 stages and normalized. It is possible to predict the material requirements using the regression model by confirming the material requirements for each process divided into 10 steps, constructing the regression equation together with the independent variables. Figure 31 shows the process briefly.

(2) Data pre-processing

The characteristics of each offshore structure and the material requirements based on the process are composed of tables, but they must be vectorized because they cannot be used directly in the regression analysis algorithm of Big Data. Figure 32 illustrates the process.



Process of Regression Analysis

Output

- Estimated material requirement by schedule
- Estimated total material requirement

Regression Analysis

Regression Eqn.1 | Regression Eqn.2 | Regression Eqn.3 | Regression Eqn.4 | Regression Eqn.5 | Regression Eqn.6 | Regression Eqn.7 | Regression Eqn.8 | Regression Eqn.9 | Regression Eqn.10

0% 50% 100%
Schedule(%)

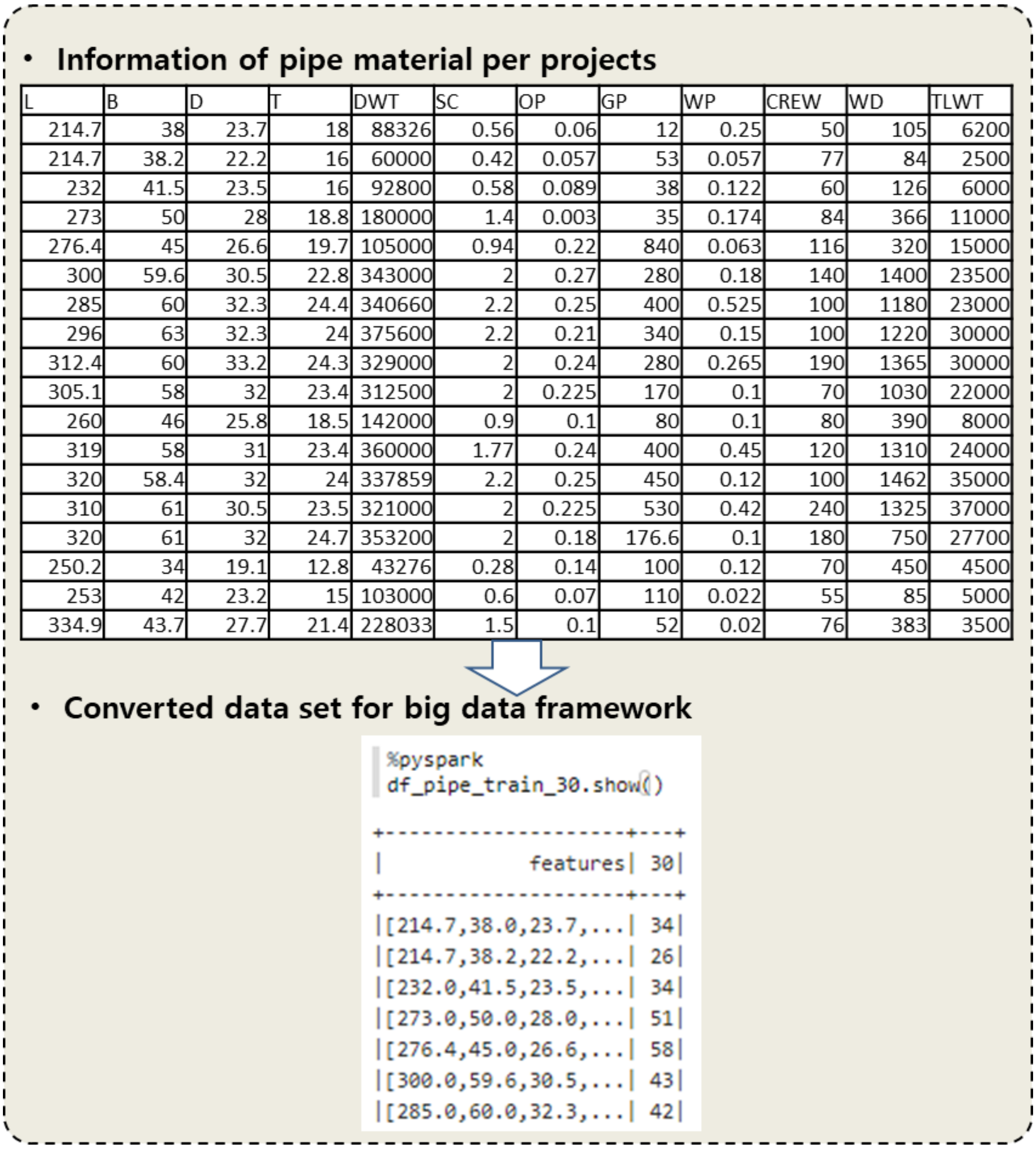
Creating equations for material requirement estimation by schedule

Input

- Project information
- Material specification
- Process rate
- Material usage by schedule

	[illegible]	LWT	Lifetime	Building	Productio	Block	Flange_To	Process	Flange_usage
A	[illegible]	7000	15	1.8	0.12	99	838	71	419
B	18037.6	20[illegible]	[illegible]	[illegible]	[illegible]35	112	928	74	464
C	20021.6	6500	17	1.9	0.1	103	735	69	480
D	17852.8	14[illegible]	[illegible]	[illegible]	[illegible]	119	869	75	435
E	19825	32000	30	3.6	0.22	90	957	70	442
F	18605	37[illegible]	20	3.7	0.22	120	984	72	492
G	19842.99	14000	15	2.3	0.14	106	978	72	489
H	14927	16000	17	2.2	0.085	108	842	65	421
I	12498.2	120[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	997	72	562
J	10176.6	4500	10	1.9	0.06	85	857	69	390

Figure 31 Overview of regression analysis



- **Information of pipe material per projects**

L	B	D	T	DWT	SC	OP	GP	WP	CREW	WD	TLWT
214.7	38	23.7	18	88326	0.56	0.06	12	0.25	50	105	6200
214.7	38.2	22.2	16	60000	0.42	0.057	53	0.057	77	84	2500
232	41.5	23.5	16	92800	0.58	0.089	38	0.122	60	126	6000
273	50	28	18.8	180000	1.4	0.003	35	0.174	84	366	11000
276.4	45	26.6	19.7	105000	0.94	0.22	840	0.063	116	320	15000
300	59.6	30.5	22.8	343000	2	0.27	280	0.18	140	1400	23500
285	60	32.3	24.4	340660	2.2	0.25	400	0.525	100	1180	23000
296	63	32.3	24	375600	2.2	0.21	340	0.15	100	1220	30000
312.4	60	33.2	24.3	329000	2	0.24	280	0.265	190	1365	30000
305.1	58	32	23.4	312500	2	0.225	170	0.1	70	1030	22000
260	46	25.8	18.5	142000	0.9	0.1	80	0.1	80	390	8000
319	58	31	23.4	360000	1.77	0.24	400	0.45	120	1310	24000
320	58.4	32	24	337859	2.2	0.25	450	0.12	100	1462	35000
310	61	30.5	23.5	321000	2	0.225	530	0.42	240	1325	37000
320	61	32	24.7	353200	2	0.18	176.6	0.1	180	750	27700
250.2	34	19.1	12.8	43276	0.28	0.14	100	0.12	70	450	4500
253	42	23.2	15	103000	0.6	0.07	110	0.022	55	85	5000
334.9	43.7	27.7	21.4	228033	1.5	0.1	52	0.02	76	383	3500

- **Converted data set for big data framework**

```
%pyspark
df_pipe_train_30.show()

+--------------------+---+
|            features| 30|
+--------------------+---+
|[214.7,38.0,23.7,...| 34|
|[214.7,38.2,22.2,...| 26|
|[232.0,41.5,23.5,...| 34|
|[273.0,50.0,28.0,...| 51|
|[276.4,45.0,26.6,...| 58|
|[300.0,59.6,30.5,...| 43|
|[285.0,60.0,32.3,...| 42|
```

Figure 32 Data pre-processing

(3) Process of regression analysis

The main characteristics and characteristics of the offshore structures are used as independent variables and are listed in Table 22.

Table 22 List if independent variables

Independent variables	
L, B, D, T	Length, Breadth, Depth, Design draft
DWT	Dead weight
SC	Storage Capacity
OP	Capacity of oil production
GP	Capacity of gas production
WP	Capacity of water production
CREW	Number of people on board
WD	Depth of well
TLWT	Total light weight

In addition, because regression models could not be constructed for all materials, we constructed a regression model for some of the commonly used materials and verified the results. Selected materials are listed in Table 23.

Table 23 List of selected materials

Material	**Material Specification**			
	Type	Size (inch)	Rating (lb.)	Material
Pipe	Bevel end	2	150	A106, A333
Gasket	Ring	2	150	S31600
Flange	Flat face	2	150	A105

Elbow	90LR	2	150	A420

Table 24 shows the coefficients of the independent variables for the flanges resulting from the regression analysis.

Table 24 Independent coefficient of flanges

	30	**40**	**50**	**60**	**70**	**80**	**90**
Intercept	-2.35491	-10.6242	5.383355	-26.3825	-24.9733	8.894943	3.720724
L	0.000988	-0.00225	0.029334	0.084208	0.043846	0.020021	0
B	0.002387	0.281861	0	0.258499	0.415164	-0.0133	0
D	0.205286	0.549592	0.454995	1.375445	1.978814	0	0
T	0.049735	-0.12802	-0.08624	-0.33576	-1.64132	-0.12983	0
DWT	0	-9.5E-07	-2E-05	-1.9E-05	-2.9E-05	-2.4E-06	-4.4E-07
SC	0	3.303434	0.672518	-10.6488	3.075268	0	-0.58476
OP	0	-10.7896	-93.302	-98.5861	9.699344	0	-18.5386
GP	0.001779	0.007572	0.005793	0	0.00162	0	0.00349
WP	0.090645	21.09809	37.07509	8.851151	-11.4474	3.597445	11.14415
CREW	0.006916	-0.04614	-0.09847	-0.06479	-0.05053	-0.02197	0
WD	0	-0.00248	-0.00531	-0.01021	-0.00898	-0.00146	0
TLWT	0	9.99E-05	0.001565	0.001994	0.000727	0.000545	0.000153
Total	0.006937	0.069499	0.179467	0.282462	0.217365	0.07842	0.028745

Table 25 shows the coefficients of the independent variables for the pipes resulting from the regression analysis.

Table 25 Independent coefficient of pipes

	30	**40**	**50**	**60**	**70**	**80**	**90**
Intercept	2.763106	-0.13688	-23.0232	66.76461	-160.252	-67.1476	0.840811
L	0.00511	0	0.566197	0.346327	0.705048	0.717023	0.102042
B	-0.02812	0.072089	-2.73168	2.488504	-1.48228	-2.19195	-1.03249
D	0.124403	1.783403	3.586186	7.608348	4.090884	1.011188	-0.61073

T	-0.07457	-0.61454	-1.26908	-14.0691	0.516898	-0.02641	0.115644
DWT	-1.7E-06	-1E-05	-5.2E-05	-4.4E-05	-0.0002	-9.6E-05	-3.2E-06
SC	0.195779	-3.29253	-24.5505	-10.7245	-5.55367	-15.6308	7.683652
OP	-4.38283	-224.865	-24.7717	-1135.05	-235.004	117.2546	-1.60437
GP	0	-0.00405	0.046101	0	-0.03374	0.00178	0.012919
WP	-0.265	-20.8476	125.4781	48.51439	13.16729	73.37606	19.01546
CREW	0	-0.04151	-0.5131	0.352579	0.016162	0	-0.05616
WD	-0.00161	-0.02177	-0.05617	0.007599	-0.05207	-0.02272	-0.0074
TLWT	0.000318	0.002898	0.008752	0.00679	0.008007	0.002289	0.001823
Total	0.008288	0.078734	0.201093	0.26761	0.210618	0.08918	0.045128

Table 26 shows the coefficients of the independent variables for the elbows resulting from the regression analysis.

Table 26 Independent coefficient of elbows

	30	**40**	**50**	**60**	**70**	**80**	**90**
Intercept	0.257623	-0.47093	8.143629	6.059641	-4.72313	-7.97326	7.583179
L	0	0	0	0.080344	0.010211	0	0.005739
B	0	0	0	-0.07603	0.026005	0.045294	0
D	0	0	0	-0.18553	0.156848	0.181472	-0.22621
T	0	0	-0.29599	-1.13062	-0.0662	0.07013	-0.11883
DWT	0	0	-1.1E-05	-4E-05	0	1.86E-06	-3.2E-06
SC	0.004405	0.164868	2.518829	5.480446	1.077122	1.825238	-0.11728
OP	0	0	-7.18311	-40.4571	23.80224	16.61222	-27.3259
GP	0	-0.00045	-0.00399	0	-0.00647	0	0.001644
WP	0	-2.85201	0	22.03516	-9.27685	-8.56748	2.889399
CREW	0	0	0	-0.01429	-0.02158	-0.0325	0.000353
WD	0	0	0.000622	0	0	-0.00194	0
TLWT	9.19E-06	0	0.000109	0.000728	0	0	0.00035
Total	0.009599	0.055307	0.185162	0.361003	0.229834	0.101977	0.034817

Table 27 shows the coefficients of the independent variables for the gaskets resulting from the regression analysis.

Table 27 Independent coefficient of gaskets

	30	40	50	60	70	80	90
Intercept	0.227524	1.47806	1.335523	-10.7364	-10.7352	-60.2235	-8.08722
L	0	0	-0.01563	-0.0061	0.016726	0.224226	0.033557
B	0	0	0.00255	0.043321	0.004448	-0.51823	-0.20417
D	0	0	0.189179	0.55973	0.188448	0	0.709573
T	0	0	0.097932	0.375867	0.471132	1.329683	0.543606
DWT	0	-4.9E-07	5.92E-07	0	0	-9.4E-05	-5.3E-05
SC	0	0	0.266544	-0.55597	-2.58916	0	0.988007
OP	0.077116	23.52354	2.032974	-1.08663	25.11834	-15.8886	2.659015
GP	0	-0.00718	-0.01232	-0.00799	-0.03124	-0.04755	-0.03746
WP	0	-9.98467	-19.8898	-15.6206	-20.3443	0	-24.2108
CREW	0	-0.0217	0.000312	-0.04711	-0.02368	-0.0233	-0.12778
WD	0	0	-0.00242	-0.0064	-0.00772	-0.01773	-0.00848
TLWT	4.91E-07	0.000182	0.000533	0.000593	0.001506	0.003187	0.002019
Total	0.009526	0.041597	0.044463	0.088238	0.175775	0.360982	0.205637

3.2.2. Prediction of material requirement and validation

The regression model can be constructed by using the coefficients of the independent variables obtained from the previous regression analysis, and it can be used to predict the material requirements again.

(1) Predicted material requirement of test project 1

Table 28 is a table showing the material requirements at the time of the process for the first test project.

Table 28 Prediction of test project 1

Type	**Schedule**							**Total usage**
	30	**40**	**50**	**60**	**70**	**80**	**90**	
Flange	15	121	276	366	260	107	50	1195
Pipe	66	583	1580	2229	1564	676	322	7019
Elbow	4	19	77	169	82	33	23	407
Gasket	4	6	8	34	66	145	74	337

A graph of the material requirements at the time of the process is shown in Figure 33. The other materials showed the highest amount in the 60 percent process section, and the gasket showed the highest requirement in the 80 process, showing a trend like the actual.

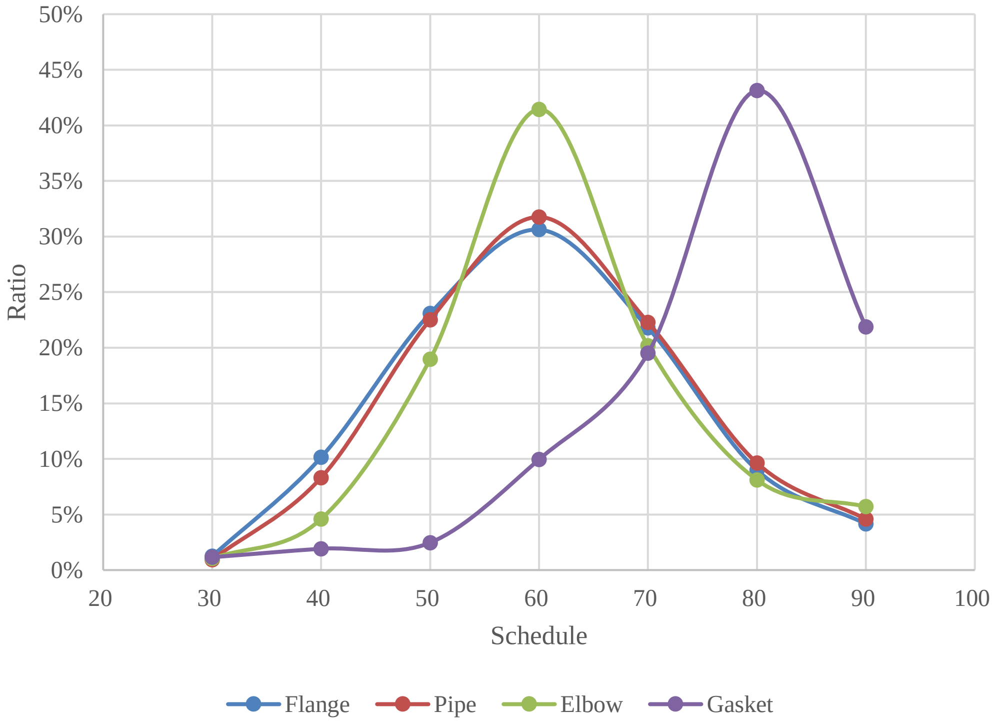


Figure 33 Trend of test project 1

Table 29 and Figure 34 compare the predicted value of the regression model and the predicted value of the artificial neural network with the actual value for the flange.

Table 29 Comparison of flange prediction for test project 1

	30	**40**	**50**	**60**	**70**	**80**	**90**
Actual	14	98	221	339	264	110	39
Linear	15	121	276	366	260	107	50
DL	6	130	213	330	305	109	36

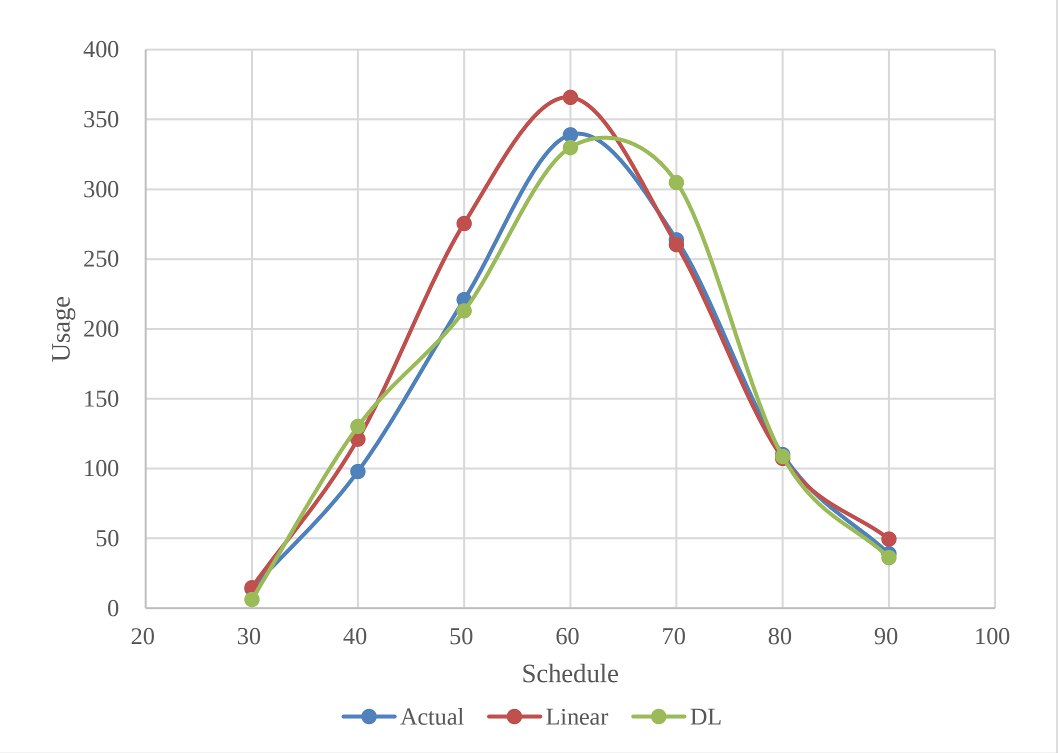


Figure 34 Trend of predicted flange for test project 1

Table 30 and Figure 35 compare the predicted value of the regression model with the predicted value of the artificial neural network for the pipe.

Table 30 Comparison of pipe prediction for test project 1

	30	**40**	**50**	**60**	**70**	**80**	**90**
Actual	67	537	1545	1947	1545	739	336
Linear	66	583	1580	2229	1564	676	322
DL	72	501	1450	2008	1394	695	267

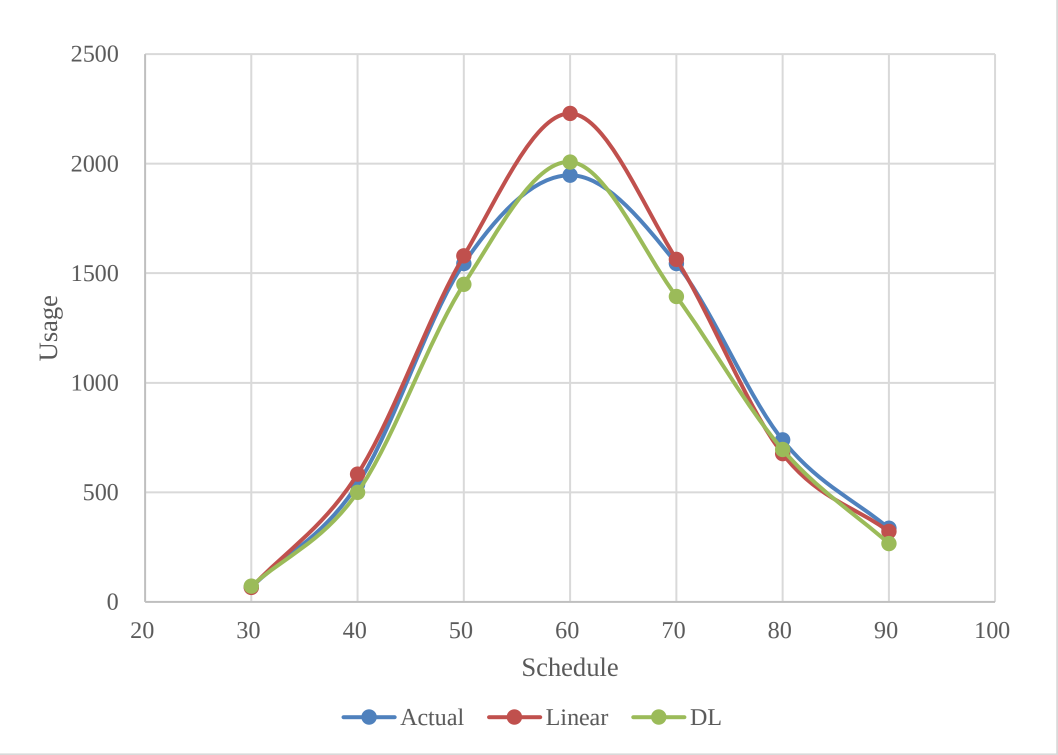


Figure 35 Trend of predicted pipe for test project 1

Table 31 and Figure 36 compare the predicted value of the regression model with the predicted value of the artificial neural network for the elbow. There are many differences in the prediction results of the neural network.

Table 31 Comparison of elbow prediction for test project 1

	30	**40**	**50**	**60**	**70**	**80**	**90**
Actual	4	20	75	154	90	36	16
Linear	4	19	77	169	82	33	23
DL	4	37	63	139	134	58	19

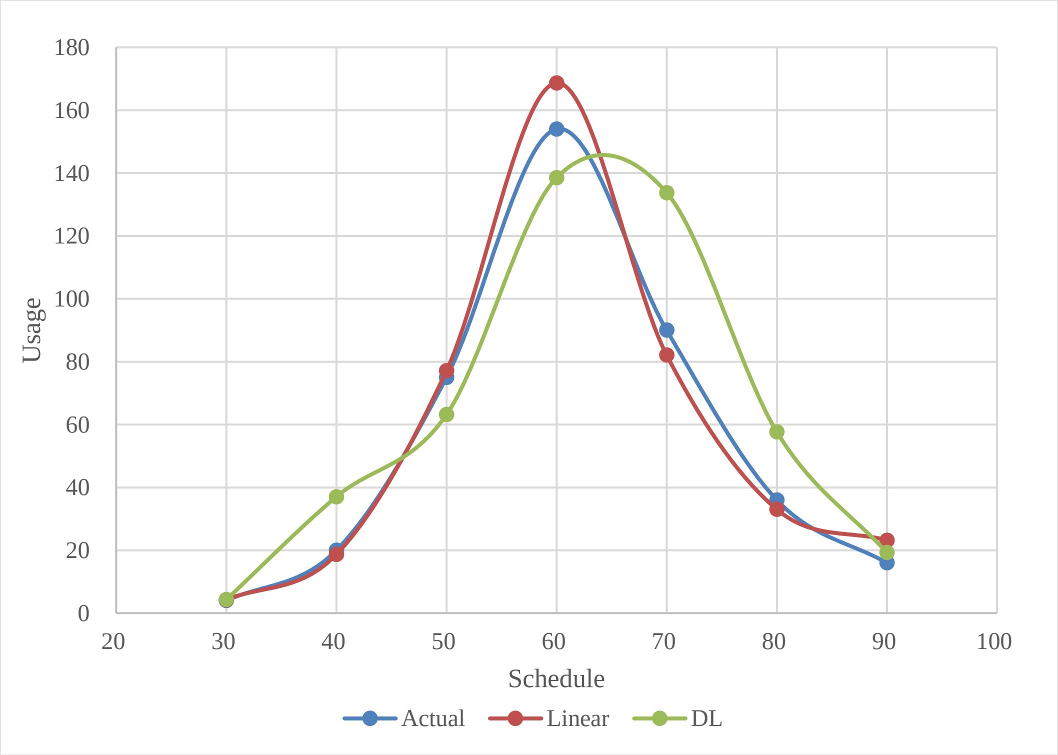


Figure 36 Trend of predicted elbow for test project 1

Table 32 and Figure 37 compare the predicted value of the regression model with the predicted value of the artificial neural network for the gasket.

Table 32 Comparison of gasket prediction for test project 1

	30	**40**	**50**	**60**	**70**	**80**	**90**
Actual	4	19	19	38	77	139	89
Linear	4	6	8	34	66	145	74
DL	11	21	31	55	86	124	119

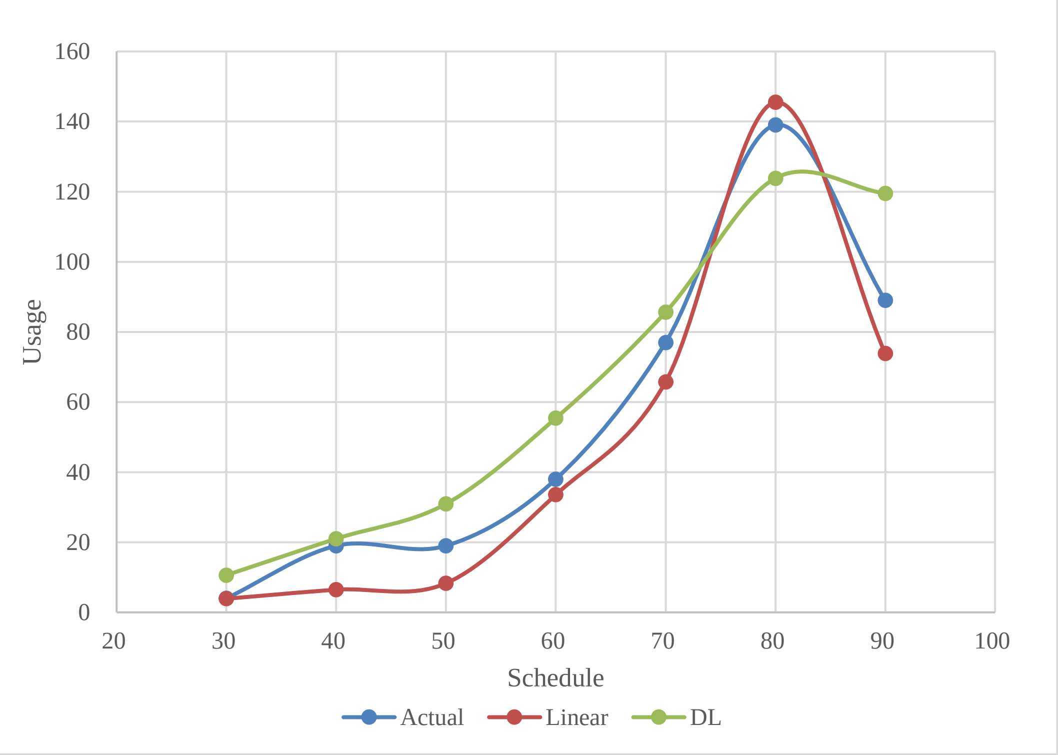


Figure 37 Trend of predicted gasket for test project 1

(2) Predicted material requirement of test project 2

Table 33 is a table showing the material requirements at the point of the process for the second test project. We can check the trend of prediction by Figure 38

Table 33 Prediction of test project 2

Type	**Schedule**							**Total usage**
	30	**40**	**50**	**60**	**70**	**80**	**90**	
Flange	13	89	211	320	253	95	34	1016
Pipe	56	509	1211	1953	1343	542	252	5867
Elbow	2	11	44	82	49	20	11	219
Gasket	5	23	32	53	103	189	120	524

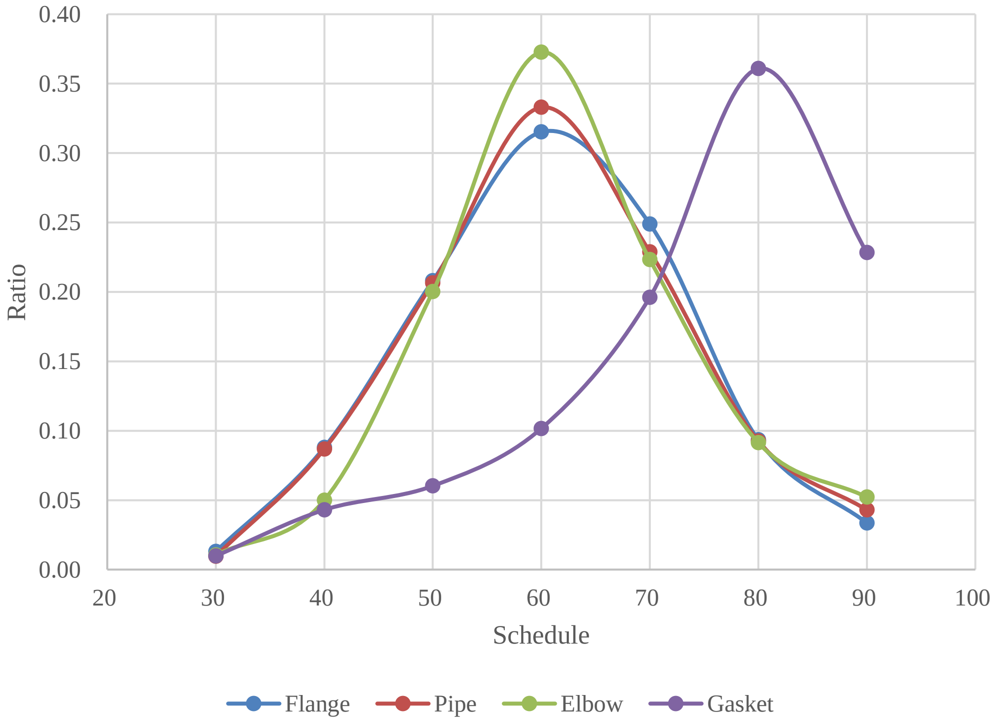


Figure 38 Trend of test project 2

Table 34 and Figure 39 compare the predicted value of the regression model with the predicted value of the artificial neural network for the flange. There are many differences in the prediction results of the neural network.

Table 34 Comparison of flange prediction for test project 2

	30	**40**	**50**	**60**	**70**	**80**	**90**
Actual	9	100	205	306	260	106	33
Linear	13	89	211	320	253	95	34
DL	4	108	199	246	247	79	31

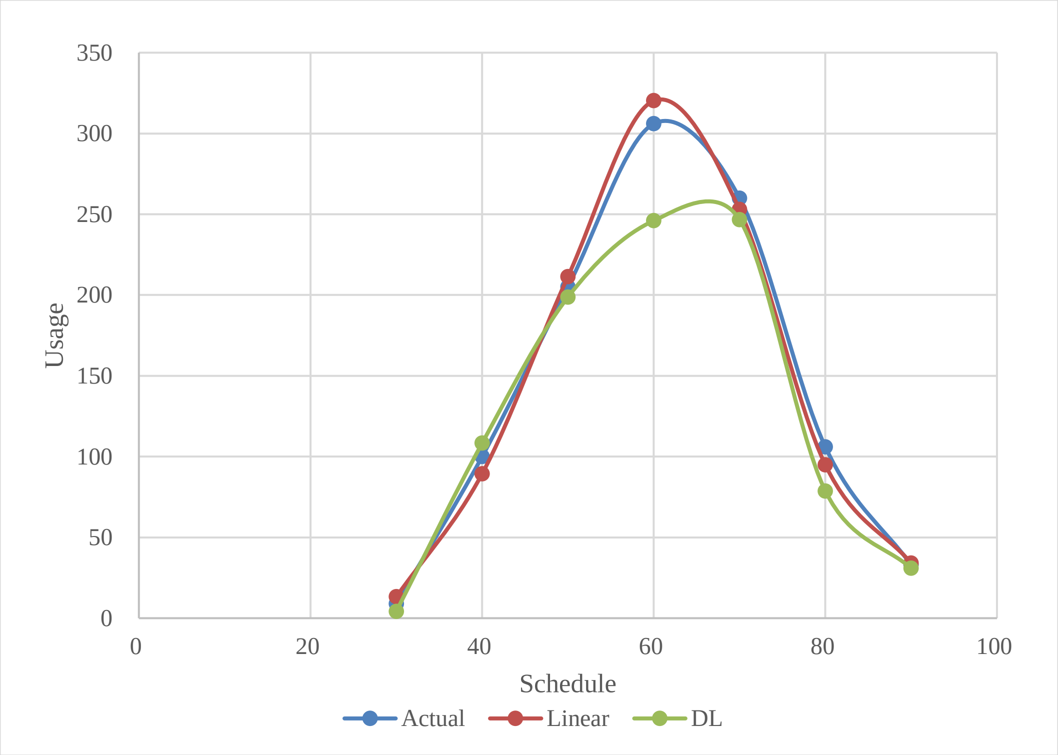


Figure 39 Trend of predicted flange for test project 2

Table 35 and Figure 40 compare the predicted value of the regression model with the predicted value of the artificial neural network for the pipe.

Table 35 Comparison of pipe prediction for test project 2

	30	**40**	**50**	**60**	**70**	**80**	**90**
Actual	59	471	1295	1884	1354	530	295
Linear	56	509	1211	1953	1343	542	252
DL	60	427	1240	1808	1253	612	227

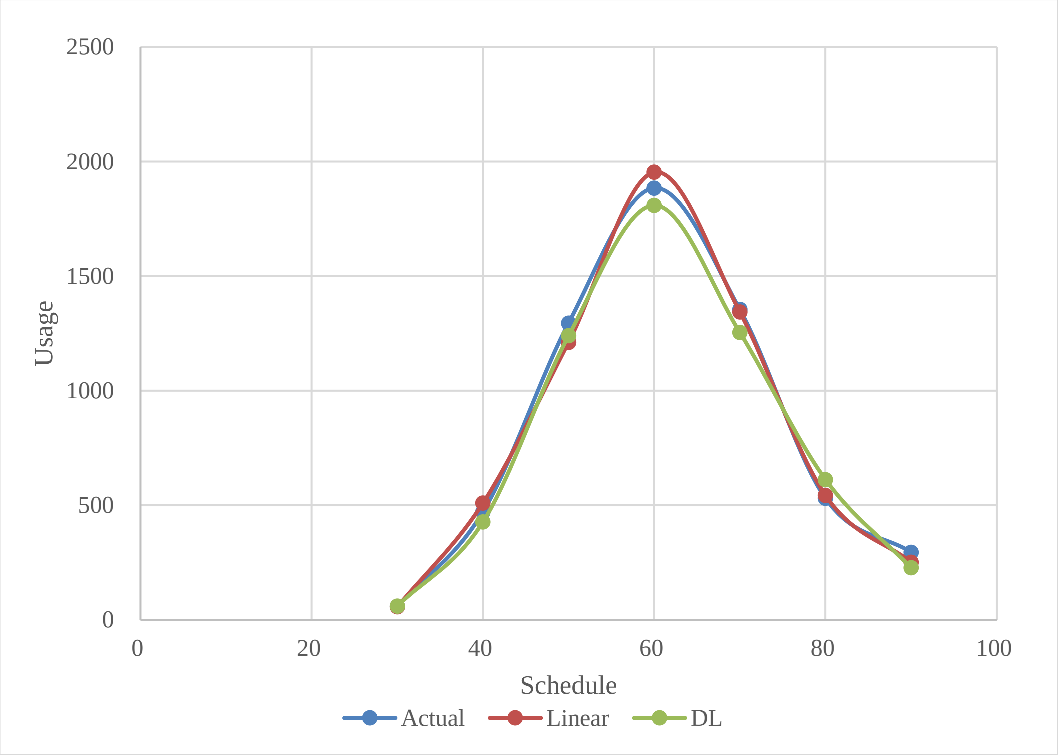


Figure 40 Trend of predicted pipe for test project 2

Table 36 and Figure 41 compare the predicted value of the regression model with the predicted value of the artificial neural network for the elbow. There are many differences in the prediction results of the neural network.

Table 36 Comparison of elbow prediction for test project 2

	30	40	50	60	70	80	90
Actual	2	10	38	75	48	20	8
Linear	2	11	44	82	49	20	11
DL	5	27	45	84	94	37	15

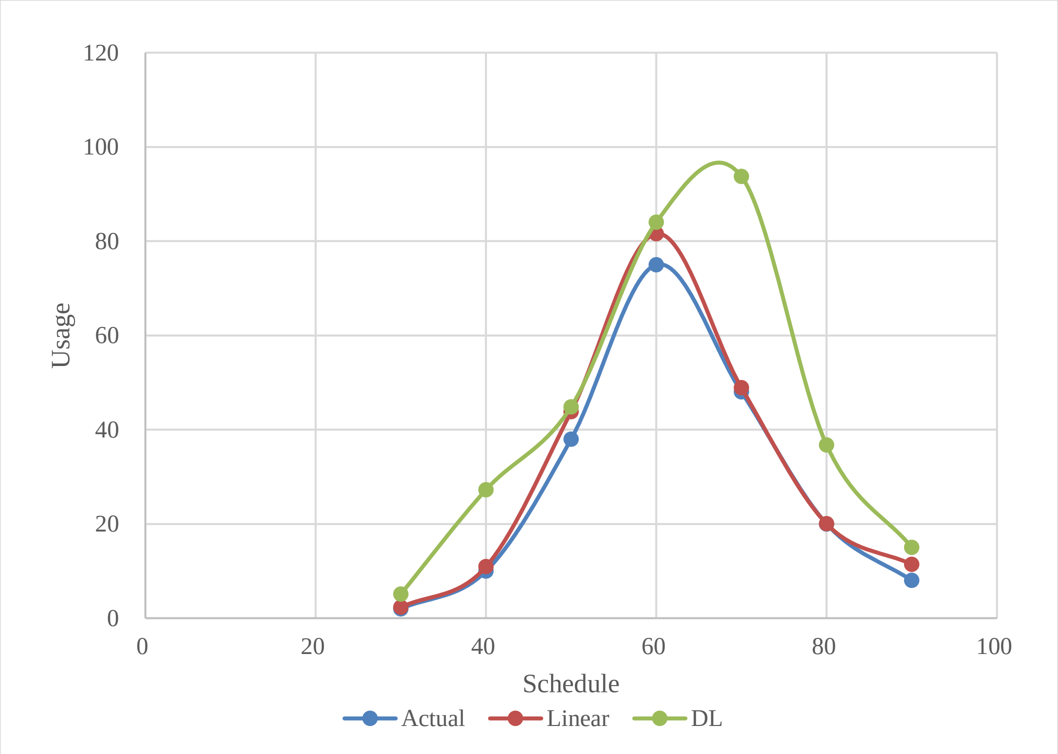


Figure 41 Trend of predicted elbow for test project 2

Table 37 and Figure 42 compare the predicted value of the regression model with the predicted value of the artificial neural network for the gasket.

Table 37 Comparison of gasket prediction for test project 2

	30	**40**	**50**	**60**	**70**	**80**	**90**
Actual	5	26	31	57	108	181	108
Linear	5	23	32	53	103	189	120
DL	4	29	31	42	96	165	141

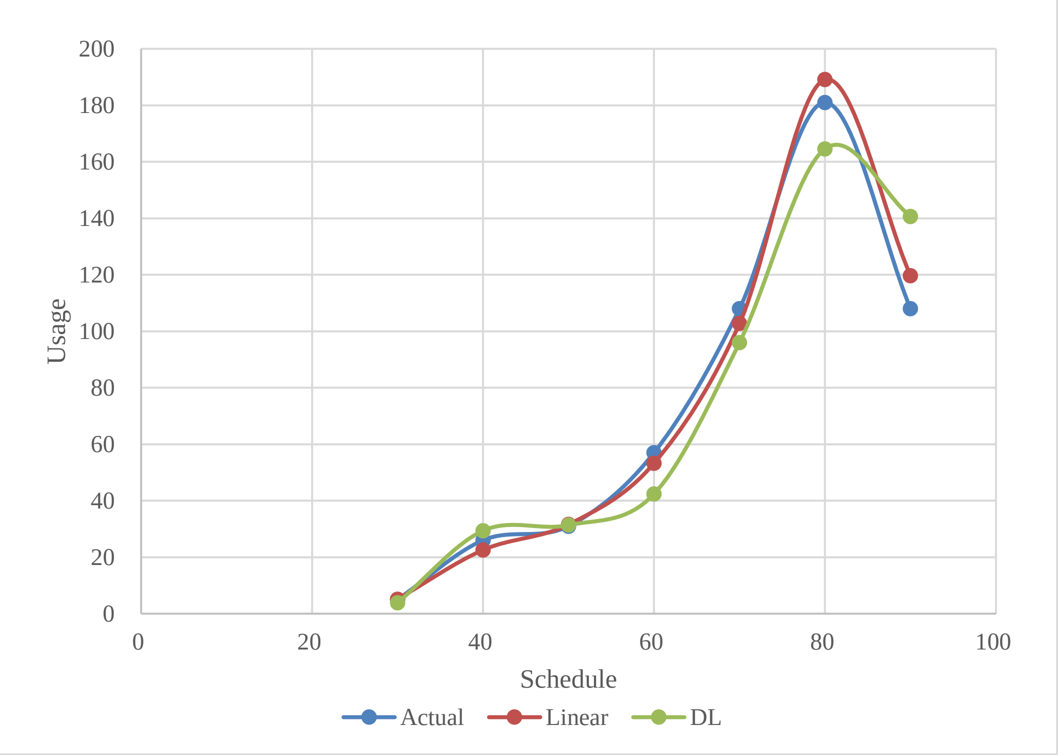


Figure 42 Trend of predicted gasket for test project 2

(3) Predicted material requirement of test project 3

Table 38 shows the material requirements at the point of time for the third test project. Figure 43 is summarizing the trend of prediction

Table 38 Prediction of test project 3

Type	**Schedule**							**Total usage**
	30	**40**	**50**	**60**	**70**	**80**	**90**	
Flange	17	114	297	457	337	131	48	1402
Pipe	81	737	1756	2679	1943	756	376	8329
Elbow	11	61	211	414	253	109	50	1110
Gasket	11	47	69	115	236	458	263	1201

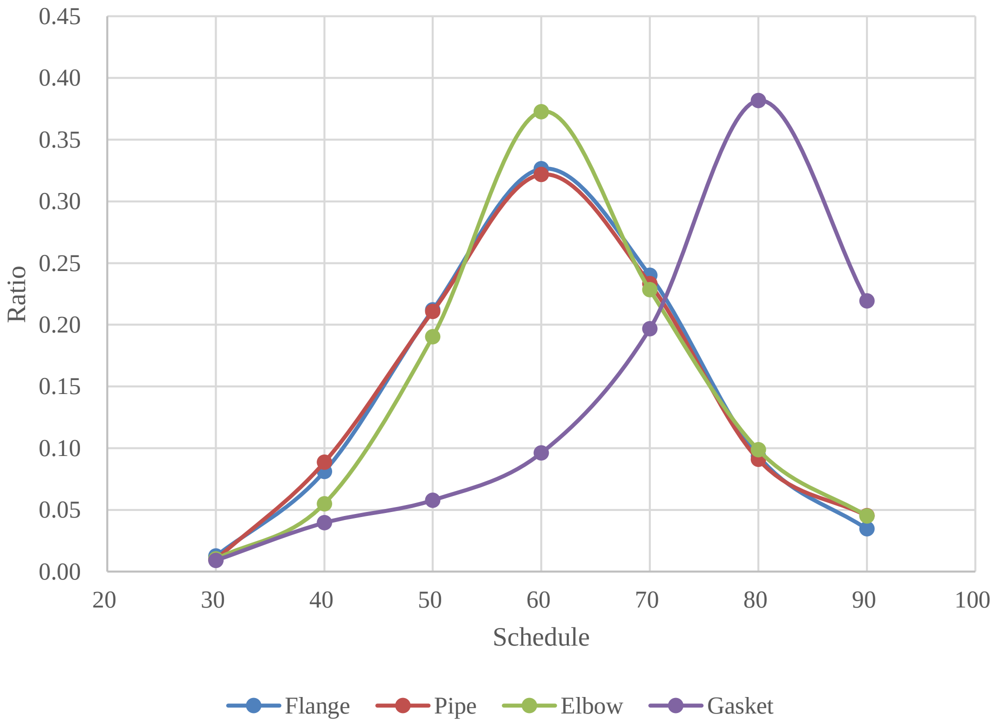


Figure 43 Trend of test project 3

Table 39 and Figure 44 compare the predicted value of the regression model with the predicted value of the artificial neural network for the flange. Both predictive models have good results

Table 39 Comparison of flange prediction for test project 3

	30	**40**	**50**	**60**	**70**	**80**	**90**
Actual	11	55	221	404	265	122	30
Linear	11	61	211	414	253	109	50
DL	13	72	199	407	286	109	46

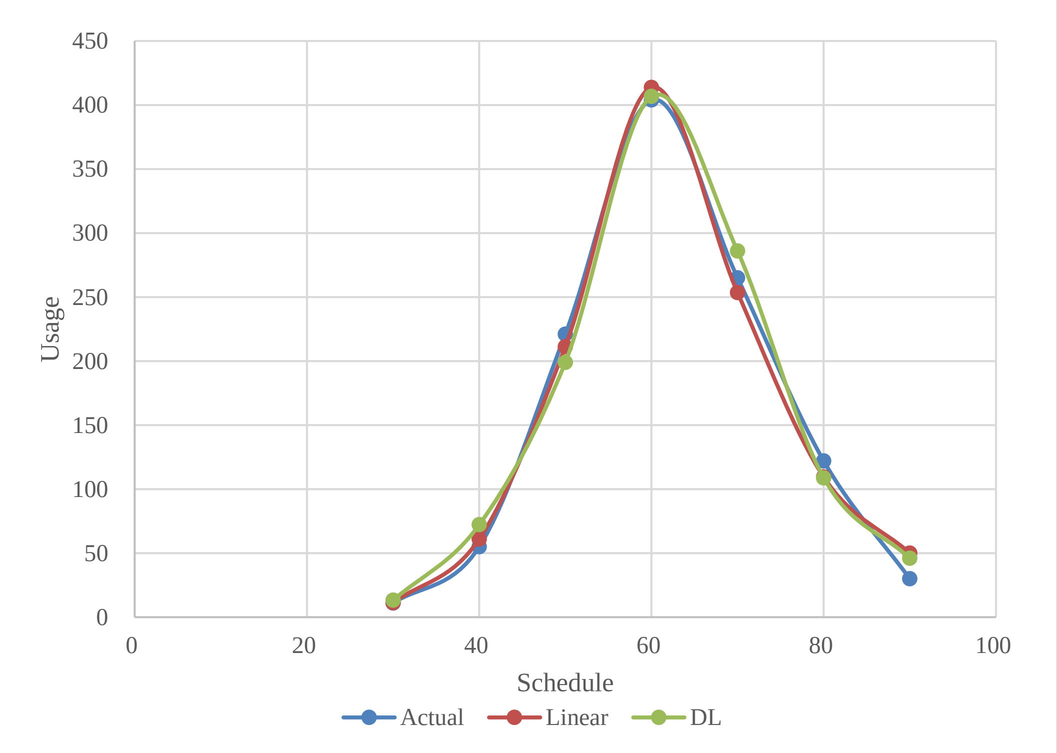


Figure 44 Trend of predicted flange for test project 3

Table 40 and Figure 45 compare the predicted value of the regression model with the predicted value of the artificial neural network for the pipe.

Table 40 Comparison of pipe prediction for test project 3

	30	40	50	60	70	80	90
Actual	24	137	277	432	327	129	45
Linear	17	114	297	457	337	131	48
DL	7	151	285	413	367	139	47

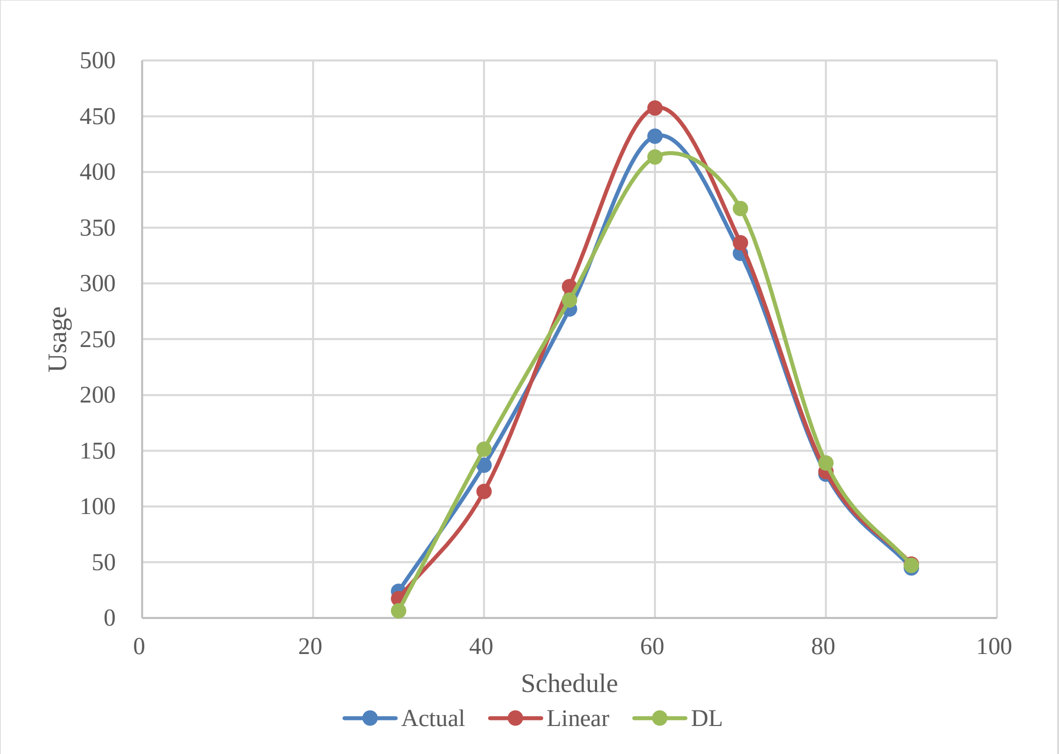


Figure 45 Trend of predicted pipe for test project 3

Table 41 and Figure 46 compare the predicted value of the regression model with the predicted value of the artificial neural network for the elbow.

Table 41 Comparison of elbow prediction for test project 3

	30	40	50	60	70	80	90
Actual	80	639	1838	2476	1838	879	240
Linear	81	737	1756	2679	1943	756	376
DL	85	624	1736	2367	1782	827	343

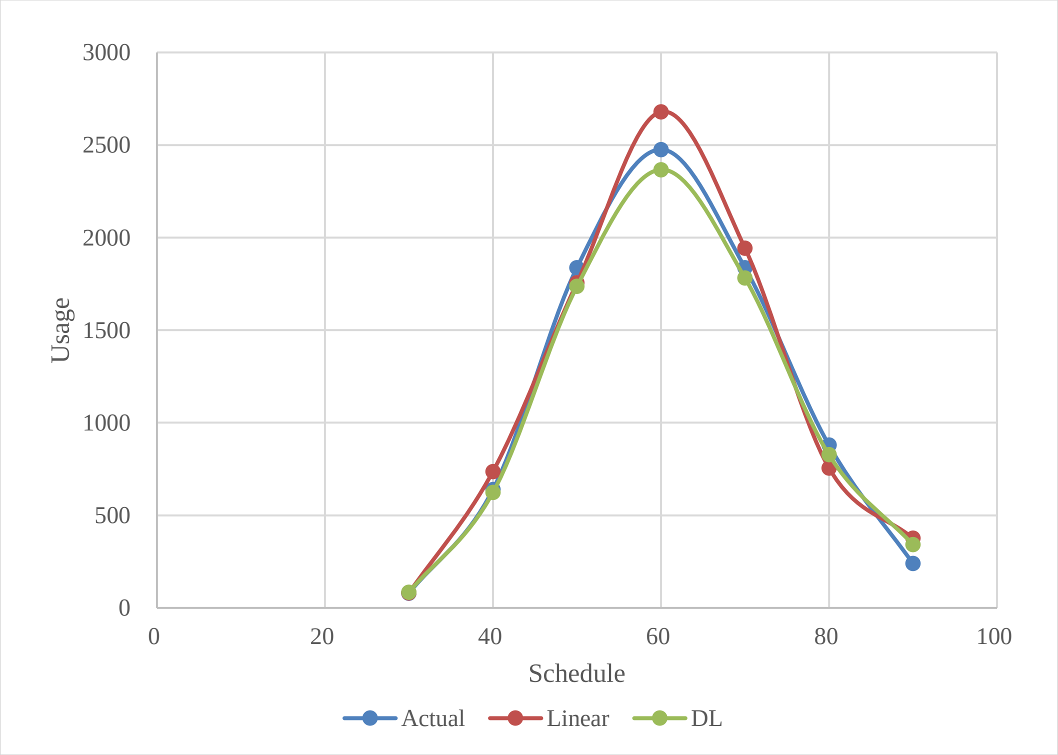


Figure 46 Trend of predicted elbow for test project 3

Table 42 and Figure 47 compare the predicted value of the regression model with the predicted value of the artificial neural network for the gasket. In most cases, the accuracy of the regression model was high, but the accuracy of the artificial neural network was high at the peak.

Table 42 Comparison of gasket prediction for test project 3

	30	**40**	**50**	**60**	**70**	**80**	**90**
Actual	10	44	68	123	233	403	233
Linear	11	47	69	115	236	458	263
DL	16	51	71	129	224	416	276

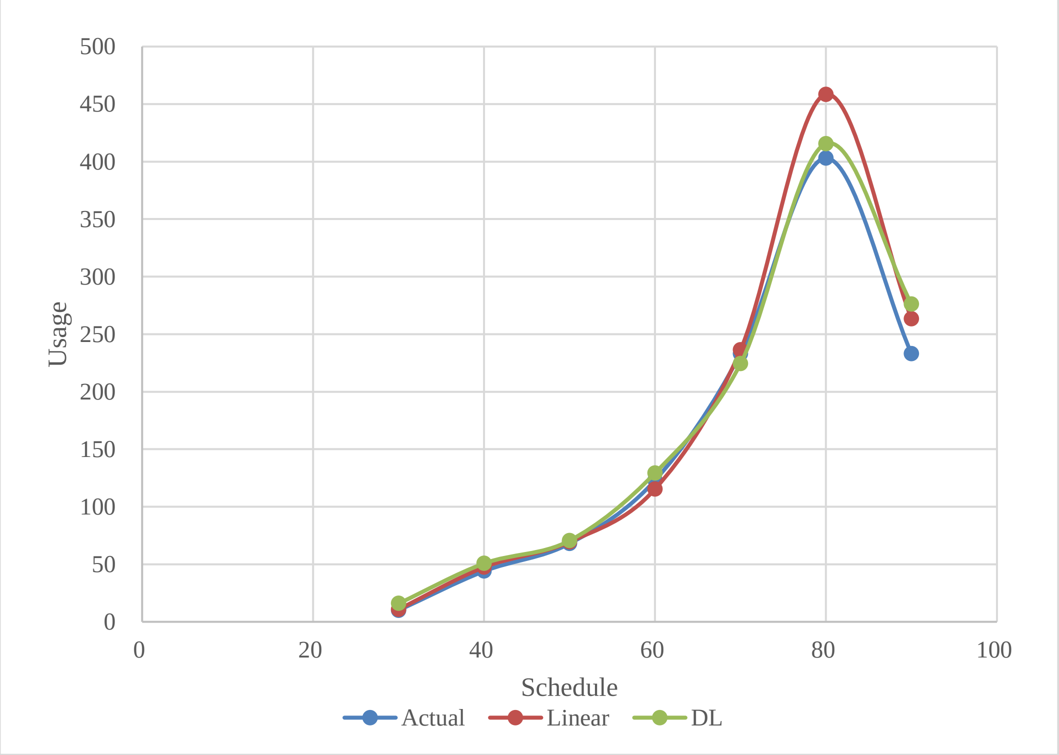


Figure 47 Trend of predicted gasket for test project 3

4. Conclusion and future works

In these days, big data technology and knowledge mining from data become important. So, this research performed to apply big data technology for shipbuilding and offshore industry. Big data technology was applied to the analysis of an offshore structure related data. Various data mining and machine learning algorithms were applied. The results all fit well with the actual design or situation. The model for prediction has shown considerable accuracy.

In association analysis, we can confirm the piping materials used together, and I think that we can make a system that recommends the material to the designer based on this. Regression analysis was able to predict material requirements very accurately and develop algorithms to build material requirements planning and procurement plans based on this.

For future works, each algorithm is going to be advanced. The machine learning algorithm will be applied to other examples

Related Works

[1] Li, J., Tao, F., Cheng, Y., & Zhao, L. (2015). Big Data in product lifecycle management. International Journal of Advanced Manufacturing Technology, 81(1–4), 667–684.

[2] Tapedia, K., & Wagh, A. (2016). Data Mining for Various Internets of Things Applications. National Conference "NCPCI, (March), 127–132.

[3] Ham, D., Lee, Y. G., & Woo, J. (2016). 조선소 의장품 조달관리를 위한 빅데이터 기반 시뮬레이션 연구. 한국경영과학회 춘계공동학술대회 논문집, 4, 3142-3149.

[4] Ham, D., Lee, P., & Woo, J. (2016). 조선소 빅데이터 활용을 위한 기계학습 적용 방법 연구. 한국 CAD/CAM 학회 논문집, 186-190.

[5] Ham, D. (2016). 조선소 의장품 조달관리를 위한 데이터 마이닝 방법론에 관한 연구. Master Thesis

[6] Saabith, A. L. S., Sundararajan, E., & Bakar, A. A. (2016). Parallel implementation of Apriori algorithms on the Hadoop-MapReduce platform - An evaluation of literature. Journal of Theoretical and Applied Information Technology, 85(3), 321–351.

[7] Zhang, Y., Ren, S., Liu, Y., Sakao, T., & Huisingh, D. (2017). A framework for Big

Data driven product lifecycle management. Journal of Cleaner Production, 159, 229–240.

[8] Li, D., Tang, H., Wang, S., & Liu, C. (2017). A big data enabled load-balancing control for smart manufacturing of Industry 4.0. Cluster Computing, 20(2), 1855–1864.

[9] Lee, Y. (2017). A Reference Model for Big Data Analysis in Shipbuilding Industry. Ulsan National Institute of Science and Technology.

[10] Kim, S., Roh, M., Kim, K., & Oh, M. (2017). Big Data Platform Based on Hadoop and Application to Weight Estimation of FPSO Topside. Journal of Advanced Research in Ocean Engineering, 3(1), 32–40.

[11] Musalem, A., Aburto, L., Bosch, M., Musalem, A., Aburto, L., & Bosch, M. (2018). Market basket analysis insights to support category management. European Journal of Marketing, 52(7/8), 1550–1573.

[12] Changhai, H., & Shenping, H. (2018). Factors correlation mining on maritime accidents database using association rule learning algorithm. Cluster Computing, 0123456789.

[13] Abbasian, N. S., Salajegheh, A., Gaspar, H., & Brett, P. O. (2018). Improving early OSV design robustness by applying 'Multivariate Big Data Analytics' on a ship's life cycle. Journal of Industrial Information Integration, (February), 0–1.

[14] Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. Expert Systems with Applications, 100, 1–16.

[15] He, Q., He, W., Song, Y., Wu, J., Yin, C., & Mou, Y. (2018). The impact of urban

growth patterns on urban vitality in newly built-up areas based on an association rules analysis using geographical ‘big data.’ Land Use Policy, 78(July), 726–738.

[16] Park, S. H., Synn, J., Kwon, O. H., & Sung, Y. (2018). Apriori-based text mining method for the advancement of the transportation management plan in expressway work zones. Journal of Supercomputing, 74(3), 1283–1298.

[17] Szymkowiak, M., Klimanek, T., & Józefowski, T. (2018). Applying Market Basket Analysis To Official Statistical Data. Econometrics, 22(1), 39–57.

[18] Oh, M.-J., Roh, M.-I., Park, S.-W., & Kim, S.-H. (2018). Estimation of Material Requirement of Piping Materials in an Offshore Structure using Big Data Analysis. Journal of the Society of Naval Architects of Korea, 55(3), 243–251.

국문 초록

빅데이터 기반 해양 구조물의 데이터 마이닝 방법

조선소에 많은 선박과 해양 구조물들이 건조되면서 다양한 데이터들이 설계 및 건조 과정에서 생성되고 누적된다. 누적된 데이터를 빠르게 처리하여 의사설정에 이용하려는 필요성에 발생함에 따라 빅데이터 기술의 도입 필요성도 함께 커지고 있다. 본 연구에서는 조선소에서 발생할 수 있는 두 가지 사례에 대하여 빅데이터 기반 데이터 마이닝 방법을 통한 해결책을 제안하고자 한다.

첫 번째 문제점은 설계 단계에서 발생할 수 있는 문제로, 설계 과정에서 적절하지 못한 자재를 선정하여 그것이 오작으로 이어지는 경우이다. 또 다른 한가지는 구매 및 조달 과정에서 발생할 수 있는 문제로, 조달 과정을 관리하기 위한 자재 관련 추가적인 정보를 검색하는데 추가적인 시수가 소요된다는 점이다. 두 가지 문제 모두 미숙련자에게서 주로 발생하며, 본 연구에서는 연관성 분석을 이용한 자재 추천과 회귀 분석을 이용한 소요량 예측이라는

데이터 마이닝 방법을 통하여 해결책을 제시하고자 한다. 이러한 시스템이 설계자의 오작률을 줄이고, 조달 담당자의 시수를 줄여줄 수 있을 것이라고 본다.

배관 정보는 그 종류가 다양하고 크기가 크기 때문에 빅데이터로 간주할 수 있다. 또한, 이런 빅데이터를 처리하기 위해서 빅데이터 기술 기반 데이터 마이닝 알고리즘이 사용되었다. 자재간 연관성을 찾기 위해서 빈발 패턴 성장 알고리즘을 이용하였으며, 자재 소요량을 추정하기 위해서 빅데이터 기술 기반 회귀 분석이 사용되었다.

연관성 분석 결과와 자재 소요량의 예측 결과는 실제 사례와 비교하여 검증하였으며, 이를 통하여 제안 방법이 설계 오작과 조달 관리 기수를 줄일 수 있는 시스템의 기반이 될 수 있음을 확인하였다.

Keywords: Big data, Data mining, Offshore structure, piping design, Association analysis, Regression analysis.

Student number: 2017-20003