APPLICATION OF DATA MINING TO FIND THE INGREDIENT IN THE MANUFACTURE OF RUBBER SOLES

PARINYA WONGCHAMPA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2013

Thesis
entitled

APPLICATION OF DATA MINING TO FIND THE INGREDIENT IN THE MANUFACTURE OF RUBBER SOLES

……
Mr. Parinya Wongchampa
Candidate

……
Lect.Waranyu Wongseree, Ph.D.
Major advisor

……
Lect. Supaporn Kiattisin, Ph.D.
Co-advisor

……
Lect. Adisorn Leelasantitham, Ph.D.
Co-advisor

……
Prof. Banchong Mahaisavariya,
M.D., Dip Thai Board of Orthopedics
Dean
Faculty of Graduate Studies
Mahidol University

……
Lect. Supaporn Kiattisin, Ph.D
Program Director
Master of Science Program in
Technology of Information System
Management
Faculty of Engineering
Mahidol University

Thesis
entitled

APPLICATION OF DATA MINING TO FIND THE INGREDIENT IN THE MANUFACTURE OF RUBBER SOLES

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science
(Technology of Information System Management)
on
May 15, 2013

……
Mr. Parinya Wongchampa
Candidate

……
Asst.Prof. Bunlur Emaruchi, Ph.D.
Chair

……
Lect. Supaporn Kiattisin, Ph.D.
Member

……
Lect. Waranyu Wongseree, Ph.D.
Member

……
Lect. Adisorn Leelasantitham, Ph.D
Member

……
Asst.Prof. Worasit Choochaiwattana, Ph.D.
Member

……
Prof. Banchong Mahaisavariya, M.D., Dip Thai Board of Orthopedics
Dean
Faculty of Graduate Studies
Mahidol University

……
Lect. Worawit Israngkul, M.S.
Dean
Faculty of Engineering
Mahidol University

ACKNOWLEDGEMENTS

The success of this thesis was due to the kind support and distribution of a number of people. I would like to firstly express my sincere gratitude to Lect. Waranyu Wongseree, as major advisor, Lect. Supaporn Kiattisin and Asst. Adisorn Leelasantitham, as co-advisors for their productive advice, consultancy, and assistance that contribute to the thesis's completion.

I am also thankful to Mr.Laksana Kittitaworn for his data support, those staffs for their facilitation in paper work, as well as TISM ' 51 friends for their mental, advisory, and documentary supports, especially Ann, Fai, Jack, etc. I am grateful to my family for their constant encouragement and aids and importantly their sponsorship for my study and research. Last but not least, my thanks and appreciation goes to those other people not mentioned here but with considerable assistance for the success of this thesis. I thank you all.

Parinya Wongchampa

APPLICATION OF DATA MINING TO FIND THE INGREDIENT IN THE MANUFACTURE OF RUBBER SOLES

PARINYA WONGCHAMPA 5136058 EGTI/M

M.Sc. (TECHNOLOGY OF INFORMATION SYSTEM MANAGEMENT)

THESIS ADVISORY COMMITTEE: WARANYU WONGSENEE, Ph.D., SUPAPORN KIATTISIN, Ph.D., ADISORN LEELASANTITHAM, Ph.D.

ABSTRACT

Shoes are made to prevent foot injuries and are the produce of different materials depending on the area. Currently, the shoe has become a fashion item and production is up in the industry, with production for both domestic and international market. The footwear industry is divided into several sub-categories, whereas, the manufacturing structure of shoes in any factory is similar. Manufacturers need to improve product quality to serve customers' demands and to reduce the cost, particularly of materials. In this study, the study focuses on rubber footwear components in which, the important raw material, in the manufacture, is rubber. Compound processes are that contain a variety of recipes cannot be fixed. Therefore data mining was used to find the ingredients needed for production.

In finding the ingredients it is necessary to take the data of the ingredients used in manufacturing for analysis. For this effort, the data was divided into 3 categories i.e. Rubber, Chemicals, and Rubber-Chemicals. Then a model was created using a Neural Network, Logistic Regression, Naive Bayes, and a Decision Tree for comparison. The comparison showed that Decision trees are accurate in the classification of the ingredients, so a decision tree was applied to this research. The technique used is J48, and SimpleCART was applied, and compared using Paired T-test in order to obtain the potential technique. It was concluded that J48 is more effective and suitable for the data.

KEY WORDS: DATA MINING / RUBBER SOLES / DICISION TREE

55 pages

การประยุกต์ใช้เหมืองข้อมูลเพื่อหาส่วนผสมในการผลิตพื้นรองเท้ายาง

APPLICATION OF DATA MINING TO FIND THE INGREDIENT IN THE MANUFACTURE OF RUBBER SOLES

ปริญญา วงศ์จำปา 5136058 EGTI/M

วท.ม. (เทคโนโลยีการจัดการระบบสนเทศ)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์: วรัญญู วงษ์เสรี, Ph.D., สุภาภรณ์ เกียรติสิน, Ph.D., อดิศร ลีลาสันติธรรม, Ph.D.

บทคัดย่อ

รองเท้าถูกทำขึ้นเพื่อใช้ป้องกันการบาดเจ็บ ทำขึ้นจากวัตถุที่แตกต่างกันตามแต่ละพื้นที่ ซึ่งในปัจจุบันรองเท้ากลายเป็นแฟชั่นด้วยจึงมีการผลิตมาขึ้นจนกลายเป็นอุสาหกรรม ผลิตขายทั้งในและต่างประเทศ อุตสาหกรรมรองเท้าแบ่งออกเป็นอุตสาหกรรมย่อยได้อีกหลายประเภท โดยที่กระบวนการในการผลิตมีลักษณะคล้ายคลึงกัน ผู้ประกอบการจึงต้องปรับปรุงสินค้าให้มีคุณภาพ ตามความต้องการของลูกค้า และลดต้นทุนการผลิตโดยส่วนสำคัญในการผลิตก็คือวัตถุดิบ ในการศึกษานี้ได้ศึกษาโรงงานผลิตส่วนประกอบรองเท้าชนิดยาง วัตถุดิบที่สำคัญในการผลิตก็คือยาง ซึ่งในกระบวนการผสมมีส่วนผสมที่หลากหลายไม่สามารถกำหนดสูตรได้ตายตัว จึงได้นำเหมืองข้อมูลมาประยุกต์ใช้เพื่อหาส่วนผสมตามความต้องการของการผลิต

การหาส่วนผสม นำข้อมูลส่วนผสมที่ใช้ในการผลิตทั้งหมดมาวิเคราะห์ และแบ่งข้อมูลวิเคราะห์เป็น 3 กรณีศึกษา คือ ข้อมูล ยาง สารเคมี และรวมยางและสารเคมี มาสร้างโมเดลโดยวิธี ข่ายงานระบบประสาท, การถดถอยโลจิสติก, เบย์อย่างง่าย และ ต้นไม้ตัดสินใจมาเปรียบเทียบกันพบว่า ต้นไม้ตัดสินใจมีความแม่นยำในการจำแนกประเภทของส่วนผสมมากที่สุด จึงได้นำ ต้นไม้ตัดสินใจ มาประยุกต์ในการวิจัยนี้ โดยเทคนิคที่ใช้คือ J48 และ Simple CART โดยนำเทคนิคทั้ง 2 นี้มาวิเคราะห์เปรียบเทียบโดย Paired T-test เพื่อหาเทคนิคที่จะนำมาใช้ ซึ่งสามารถสรุปได้ว่าใช้ J48 ให้ผลที่ดีกว่าและเหมาะสมกับข้อมูล

55 หน้า

CONTENTS

CONTENTS (cont.)

CONTENTS (cont.)

LIST OF TABLES

LIST OF FIGURES

CHAPTER I
INTRODUCTION

1.1 Background and Origin of the Problem

Shoes are daily used footwear created by human. Shoes created in different areas are generally dependents on production factors of local climate and culture and hence varied by their materials. Shoes are intended to protect foot and prevent it from pain caused by walking or running. There are a lot of categories of shoes with different materials and functions. To make a pair of shoes is to stitch up and assemble the component pieces. A variety of materials are used depending on manufacturers which include such as plastics, wood, fabrics, leather, rubber, etc.

Footwear and accessories industry [17] is, among others, provides significant contribution to national economy as a source of income and foreign-currency earnings. Footwear is a labor-intensive industry. The manufacturing structure of footwear and accessories industry is not much complicated however requires much of labors and labor skills. Many of manufacturing procedure are not possible to apply technologies because shoes are subject to changing styles and fashions, making it difficult to base on machine. For these reasons and limitations, manufacturing structure is more or less the same for this industry, and the cost advantage and disadvantage are dependent on a number of major fundamental factors including labor skill, product design proficiency, product quality development efficiency, wages, ability to acquire or availability of inexpensive sources of raw materials, availability of supplies and equipment, and operating cost, which differ from site to site.

Footwear and accessories industry can be broken down into many lines of manufacture including those of rubber and plastic shoes, athletic shoes, sandal, genuine and artificial leather shoes, and other types of shoes, as well as parts of shoes. Footwear and accessories industry involves 3 modes of production, namely, production to order of oversea parent company, production to order or subcontracting

to well-known oversea manufacturer under the trademark, specific design and material, production by domestic manufacturer and local design. Footwear and accessories production is dominantly small and medium sized industry.

The site investigated by the researcher in the current study is a footwear parts and accessories factory with specific production of shoe sole. The factory itself designs shoes for production and distribution. There are altogether 1,000 designs, each of which is used to produce 3 types of shoes, i.e. PU, PVC, and rubber, depending on customers' requirement, however, the production focuses on rubber shoe soles because of having top order, followed by those of PVC and PU. The factory is capable of producing 3,000 pairs of rubber shoe sole per day.

Rubber is one among materials used for producing parts of shoes, depending on the type and quality of rubber. To produce a rubber-based product, it requires a proper choice of type of rubber and chemicals to aid in processing. As each chemical differs in its property and action, the chemical-rubber mixture needs to meet suitable proportion to provide intended quality. Normally, there is no fix formula for any types of rubber products; it is thus helpful to modify the formula to offer more varieties of products. Shaping rubber-based products also takes several modes based on the type of particular product. To obtain the desired and different qualities of product, different formulas and processes have to be chosen.

The process of making each type of shoes is similar which involves developing rubber compound, shaping, cutting, and coloring. The crucial procedure is to prepare the rubber-chemical mixture according to the calculated proportion. The calculation is made by prescribing the quantity of each ingredient as an amount based on a total of 100 parts of the rubber which is called phr or pphr, and hence the compounding formula is obtained.

With high competition of both domestic and global arenas, factories are required to improve their product quality to meet customers' requirement and reduce cost. Importantly, materials particularly rubber is not stable in price. Today, EXCEL program is used by factories to calculate and acquire the ratios of ingredients and this has made it difficult to modify the mixture with multiple formulations at a time as errors might occur. Moreover, each mixture contains different ingredients each of which possess different properties, and rubber compounds can be varied enormously

to be chosen from depending on their properties such as color, price, softness, elasticity, etc. Besides, there are chemicals used in order to add other properties, and that the formulations differ and not possible to be fixed. The production requires those with expertise or experience to calculate, improve, and develop formulas to meet the requirement before proceeding to shaping the product. However, very limited number of people are proficient in formula development and as a result rubber production are facing with problems and disadvantages in this highly competitive business. Further, training expert for this job will be very costly in terms of time and money.

Therefore, the researcher is interested to apply data mining to assist in calculating the ingredients of rubber compound for producing shoe soles. This will also help reduce errors caused by compound modification and the cost for formula testing, as well as skill training or hiring experts. Information on individual formula developed by the factory is to be analyzed for its composition used in rubber compounding in order to meet the requirement.

1.2 Objective

1. Collect the data to find the ingredients of rubber and chemical to suit the hardness of shoes.

2. To use the result for analysis and identifying model of rubber compound.

1.3 Scope of study

1. The data employed is the rubber compound used in actual process to produce shoe soles by the factory.

2. The current study does not include chemical reaction as variable in the calculation since the data obtained is based on actual rubber compounding by the factory, and the values obtained are used by the researcher for further analysis.

1.4 Expected Benefits

1. The test results help to acquire what kind and quantity of chemicals used in rubber compound in order to meet particular requirement.

2. The test results will enable cost reduction and increased production efficiency.

CHAPTER II
THEORY

2.1 Rubber and Types of Rubber

Rubber [1] is a natural polymer with a number of properties incomparable by other materials particularly in its elasticity. When applying tensile force, rubber can stretch by many folds of its length, and after releasing the force, rubber will return to its original length and shape. Other prominent properties include toughness and high level of abrasion resistance, waterproof and airproof, high adhesive bonding with other materials. Accordingly, rubber offers multiple applications. Each type of rubber products differs in functional condition; some require oil resistance, while others need tolerance to deterioration. Suitable type of rubber should be selected to meet functional condition of the product.

Synthetic rubber refers to not only artificial rubber with physical and chemical properties similar to natural rubber, but also various types of rubber synthesized from chemical reaction. It properties include high elasticity, power absorption and drainage, compression set, good strength, tear strength and flex-cracking resistance, and abrasion resistance. It exhibits heat and electrical insulation, and chemical resistance. Today, both natural and synthetic rubber are commonly used which are categorized into 3 groups [2] shown below.

Table 2.1 Group 1: Rubbers used for general purposes, less expensive and high volume of use

Rubber	Usage properties	Functional Example
Natural rubber	Strength, abrasion resistance, good elasticity	Aircraft tires, rubber fencer, rubber bridge bearing

Table 2.1 Group 1: Rubbers used for general purposes, less expensive and high volume of use (cont.)

Rubber	Usage properties	Functional Example
Isoprene rubber (Polyisoprene)	Similar to natural rubber	Substitute to natural rubber
SBR rubber (Styrene-butadiene)	Similar to natural rubber but better resistant to abrasion and deterioration	Car tires
Butadiene (Polybutadiene)	Better elasticity and abrasion resistance than natural rubber, good function in low temperature	Car tires
Butyl rubber (Isobutylene-isoprene)	Low permeability to gases	Tire inner liners
EPDM (Ethylene-propylene)	Oxidation and sunlight resistances	Windshield rubber, window seals

Table 2.2 Group 2: Rubbers with unique properties, more expensive, less volume of use

Rubber	Usage properties	Functional Example
Nitrile rubber (acrylonitril-butadiene copolymer)	Oil resistance	Rubber ring, rubber tube
Chloroprene (Neoprene) rubber (polychloroprene)	Ozone and oil resistances	Rubber liners, engine belt, diving suit
Fluorocarbon rubber	Very high temperature resistance, chemical and oil resistance (very expensive)	Engine compartment, aircraft and spacecraft compartments

Table 2.2 Group 2: Rubbers with unique properties, more expensive, less volume of use (cont.)

Rubber	Usage properties	Functional Example
Silicone rubber	Low and high temperature resistances	
Urethane rubber (polyurethane rubber)	Very high abrasion resistance Use as foam	Printing roller, car belt and accessories, furniture seating
Chlorosulphonated polyethylene (Hypalon) rubber	Good ozone resistance Very good electrical insulation	Rubber liners, rubber cover

Table 2.3 Group 3 of rubbers with other unique properties similar to those in group 2, with somewhat better properties but more expensive and hence low volume of use

Rubber	Usage properties
Epichlorhydrin rubber	Heat resistance, Very low permeability to gases, oil and ozone resistance
polysulphide rubber	Oil and chemical resistance at high temperature
Ethylene acrylic rubber	Heat and oil resistance
EVA rubber (ethylene-vinyl acetate copolymer)	Ozone and air resistance

2.2 Rubber Formulation

Rubber formulation [1] is crucial for the quality and cost of product. Rubber formulation requires extensive knowledge on rubber specification, function, and chemicals to be mixed with rubber. The price of chemicals used should also be considered whether they are suitable or cost-benefit for producing particular products. The base of compounding formula contains the following groups of ingredients.

2.2.1 Rubber

Rubber formulation requires good knowledge on the specification of each type of rubber, as well as advantages and disadvantages. Today, technologies are employed for compounding natural and synthetic rubbers in order to yield the product containing good properties of rubber of each type.

2.2.2 Vulcanizing agent or curing agent

This group of substances effect the change of rubber properties to attain high elasticity or so called "vulcanization", and it is referred to as "vulcanized rubber" by the factory. There are 2 major groups of substances used for vulcanization process, i.e. sulphur system and peroxide system. Others include metallic oxide such as magnesium oxide (MgO) and zinc oxide (ZnO).

2.2.3 Antidegradants

Since natural and synthetic rubbers are vulnerable to certain factors such as sunlight and oxygen that affect rubber degradation. Filling of antidegradants is thus necessary to prolong the use life of product. Examples of antidegradants are IPPD TMQ BHT.

2.2.4 Fillers

Fillers serve to reinforce rubber product and reduce cost. They include such as carbon black and silica. Non-reinforcing filler preferably used to reduce cost are clay, talcum, calcium carbonate, etc.

2.2.5 Processing Aids

Substances in this group differs in individual unique action. Softening agents used during the mixing process are oil and peptizer. Some agents aid in controlling the nerve of rubber such as factice.

2.2.6 Miscellaneous Ingredients

In general this group of agents is not required in rubber formulation, however in case certain special rubber properties are required, a filler is needed. For

example, a retarder is used to prolong rubber shaping during the process, blowing agent is used to generate gas and expand rubber in making sponges, or agents that produce colors.

In rubber formulation, the quantity of different substances will be identified in proportion to 100 parts of rubber by weight called as phr or pphr (part per hundred of rubber).

Table 2.4 Functions and amount of different agents used [14]

Required properties	**Chemical agents**	**Amount used (phr)**
Elasticity	Vulcanizing crosslink agents Activator Accelerator	1-3.5 1-5 0.5-2.5
Antedegradants for O_2 , O_3	Antedegradants e.g., 6 PPD, Flectol H, Antioxidant 2246, Wingstay L, Vulkanox MB	1-4
Rubber reinforcement	Small particle fillers, e.g. carbon black, Silica	10-100
Reduced viscosity of raw rubber	Mastication and/or chemical peptizer	0.1-1.5
Adhesion with fabrics, metal, tiles	Bonding agent and requiring cleaning or special method to treat the surface of object to be bonded	2-10
Color	Organic or inorganic colors	Per required intensity
Cost reduction	Inexpensive fillers, reclaimed rubber, scraped rubber	10-200 , 10-100, 5-50
Cellular structure	Blowing agent, organic or inorganic agents	05.-20, 5-30

Table 2.4 Functions and amount of different agents used [14] (cont.)

Required properties	Chemical agents	Amount used (phr)
Self-extinguishing	Flame retarder, e.g., phosphates, antimony salts, halogenated organics, borates (antimonytrioxide and chlorinated wax – often used with natural rubber)	1-20
Electrical insulation	Non-conductive agents, e.g., fillers such as mineral, oil, hydrocarbon	5-50
Antistatic	Antistatic agents, e.g., polarized ester, carbon black	0.1-2.0, 1-5
Electrical conduction	Conductive agents, e.g., carbon black, metallic particles metal particles, and metallic salts	10-50
Antibacteria	Antifungal agents, e.g. Chlorinated phenol	0.5-5.0

Table 2.5 Ingredients and amount of rubber and chemicals in basic formula of rubber products [14]

Ingredients	Amount (phr)
Rubber (one or two types or more)	100
Sulfur	2.5-3.5
Activator	1-5
Accelerator (one or two types or more)	0.5-2.0
Filler	(as needed)
Plasticizer, peptizer	5-10
Antidegradant	1-2

The proportion of substances used [14] depends on several factors such as the type and efficacy of chemicals, molding process, product quality. Suitable rubber formulation requires adequate knowledge and understanding on the properties, functions, and amount of both the rubbers and chemicals. However, merely proper

rubber formula does not necessarily mean that good quality product is yielded since there remains many other factors to be investigated and comprehended.

2.3 Rubber and Chemical Mixing

A machine is used for mixing rubber [1]. There are 2 types of mixer generally used based on functional mode described next.

2.3.1 Open mill

Two-Roll Mill is widely used in rubber industry from the past. It consists of 2 pairs of grinding rollers which turn toward each other but at different speeds, depending on the design. Mostly, the slower is at the front on the operator's side and the faster is at the rear. In some case however, to ease the operation, it is designed to allow the fast roll at the front because rubber of particular type especially the artificial one often bind to the roller when operating with high speed. The design of fast roll at the front helps improve processing.

Disadvantages of the two-roll mill

- The mixing cycles take long time.
- Efficiency is dependent on the operator's skill
- High levels of dust and dirt
- Difficult in controlling the uniformity of different batch

Today it is preferable to use the two-roll mill for mixing chemical agents to the masterbatch for vulcanization rather than mixing different ingredients alone.

Figure 2.1 Two Roll Mill

2.3.2 Internal mixer

Internal mixer is of large size with capability of breaking down and mixing rubber and chemicals in larger amount than the open mill. It works efficiently and fast. For internal mixer, mixing occurs in mixing chamber which is top loaded with pressure exerted on the batch using the ram. The 2 rotors are in parallel each with blades on its surface to push rubber and chemicals into mixing. Internal mixer is suitable and required for a large amount of mixture particularly for mixing ingredients with carbon black for it helps reduce the spread of carbon black.

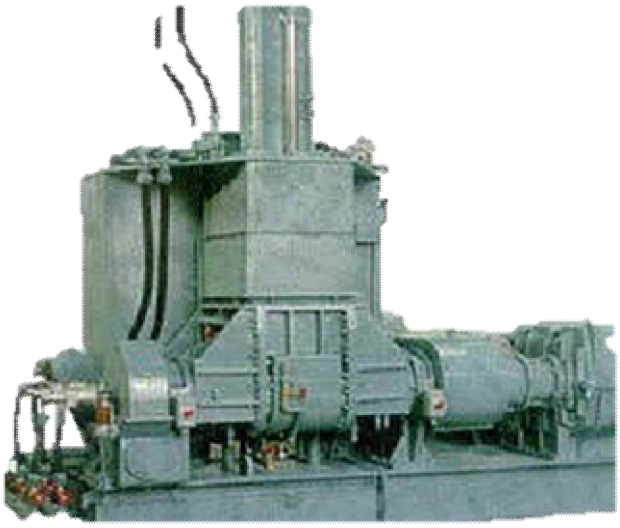

Figure 2.2 Internal Mixer

The process of breaking down rubber is technically called mastication. It is a very important procedure in rubber mixing process particularly for the one with high viscosity such as natural rubber and consequently with very high molecular weight. Mastication reduces rubber's viscosity or soften rubber to be able to mix with chemicals, allows compounding ingredients to be homogeneously distributed into the rubber. However for synthetic rubber with low molecular weight, mastication is less effective and it is possible to choose synthetic rubber with required viscosity and molecular weight during this process. Rubber molecule is crushed by means of mechanical shearing, resulting in a reduced molecular weight. The efficacy of mastication depends on temperatures including the type and amount of peptizer. At low temperature, rubber's viscosity is relatively high, hence high shear stress in the mixer, resulting in masticated molecule by mechanical shearing. With higher temperature, the rubber's viscosity reduces, hence reduced shear stress and reduced efficacy of mastication by mechanical shearing. However, in case peptizer is added, the mastication efficiency will increase as temperature increases because peptizer works well in high temperature. The reduced viscosity of rubber as a result of mechanical shearing and peptizer will be permanent. However, by heat or increased temperature, reduced viscosity can be temporary or permanent depending on the type of rubber. Rubber mastication is a very important process in rubber processing. It brings the raw rubber into changes of both physical and chemical properties. The process starts from softening rubber, then adding chemicals to enable further processing and obtain required properties. Rubber mixed with chemicals is referred to as rubber compound. Rubber compound added with only one or two chemical substances is called masterbatch. To give an example, 50% sulphur masterbatch means raw rubber mixed with sulphur 50 parts in 100 parts of compound.

Mixing chemicals with rubber [1] comprises 3 steps, i.e. communiation, incorporation, and dispersion.

Step 1: This is called pre-mixing process by which rubber is crushed and softened to reduce molecular mass by using mechanical energy from Two Roll Mill, and become plasticity with viscosity and glide.

Step 2 : This is a process occurs during chemicals are added and mixed with rubber (Incorporation). Chemical powders are impacted to rubber until totally

incorporated into rubber and the batch size reduces because the air in chemicals is expulsed. The density of rubber compound become increased. Such increased density is an indicator of efficiency in rubber compounding

Step 3: This process involves the dispersion of chemical fillers into rubber matrix. At the same time chemical agglomerate will be broken down into small particles of less than 0.1 micron in size. Chemicals disperse more into rubber in the area with high shear stress.

For the efficiency and uniformed quality of rubber mixing, it is necessary to follow proper sequences of adding chemical fillers. The general principle is that after the rubber is masticated and soften, the fillers difficult to disperse in rubber matrix will be applied first which include such as ZnO, stearic acid, carbon black powder as the temperature is still low and rubber has high viscosity, and mechanical action is high. After that, non-reinforcing fillers, other substances, and oil will be added. The fillers recommended to be added last are accelerator, sulphur, and scorch protective agents. The end product of rubber after the addition of fillers is called rubber compound.

From the above mentioned processes, methods and materials used in manufacturing, but since this study aimed to use data mining to find the ingredients, in which the important to note about Data Mining, Methods, Processes and Techniques which will be mentioned to this

2.4 Data Mining

The definition of data mining

There are a number of definitions given for data mining as follows.

Data mining [2] is the process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions,

Data mining [3],[4] is the process of sifting through large databases for hidden information in order to predict trends and behaviors based on past information, and to utilize such information to support business decision. Information obtained

can be used to create prediction or model for classification of units or clusters, or to demonstrate the relationship between different units, or to provide conclusion of the essences in databases. Data mining is part of a larger process called Knowledge Discovery in Database : KDD. Acquiring the trends of hidden data and information in large databases is important and if such information is not utilized, it will be kept useless. Data mining is a technique of discovering hidden predictive information from large databases particularly data warehouse. Data mining enables the potential in the use of information in databases.

Data mining [5],[6] brings together techniques from various works such as recognition, machine learning, statistics, and database with an aim to find the patterns of relationship among data, and analyze for useful information underlying large databases in order to obtain general actionable and useful information. Most of data mining takes into account and emphasizes on the mining steps or searching for the direction of data. In fact mining information is only part of data mining process, and data mining is part of the process called Knowledge Discovery in Database (KDD).

According to these definitions, it can be concluded that data mining is a process, technique, and method used for extracting information for further utilization. Data mining is only part of the Knowledge Discovery in Database

Knowledge Discovery in Database:KDD [7] refers to the process the broad process of finding knowledge or information in data in a large set of data. Data mining procedure is an important process in search of characteristics of interests of the data such as the patterns of relationship, prominent structural change or abnormality of data from a large amount of data stored in databases. Multiple methods used to mining data differ in their objectives depending on the overall desired results of the process. Therefore, a variety of methods should be suggested for diverse goals to offer desired results after their application.

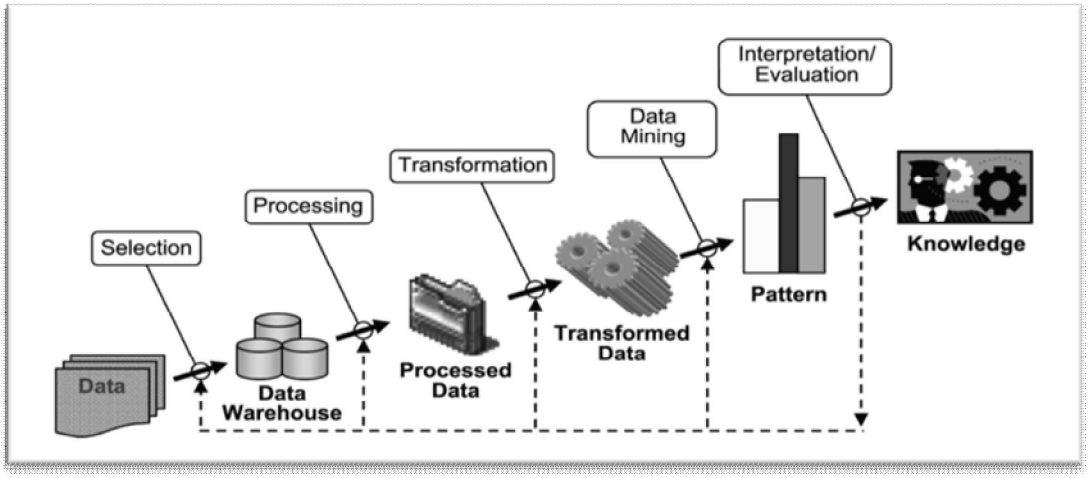

Figure 2.3 Steps of Knowledge Discovery in Database Process

KDD process [8] comprises 4 key stages.

1) Business Object is the first stage of the process which involves identifying clear scope, goals and objectives of KDD or data mining before proceeding to data mining. It includes preliminary analysis of what data will be used for and what is to be decided.

2) Data Preparation is the step of preparing and managing appropriate information as inputs to data mining process. This stage consists of 3 sub-processes.

2.1) Data Selection involves the selection of data relevant to intended goals and objectives so that the result of analysis can be used to facilitate decision making as required. Multiple data sources are available such as data warehouse, database, etc.

2.2) Data Preprocessing is a step of making adjustment of selected data for better quality since they are derived from multiple sources and hence with different formats of data storage that may result in potential errors in data mining process. This step of data preprocessing includes the following processes.

- Data Cleaning involves filling missing data fields; selecting necessary information for data mining process such as relevant data for analysis of customer's demand; deleting unnecessary details such as identification number, phone number; handling outliers and inconsistent data such as age versus year of birth, dates belong to the 13^{th} month, female with the title "Mr.", etc.

- Data Integration is to compile relevant data and examine whether the data collected from several sources are stored in similar or different formats since those with different formats may vary in measurement unit or symbols. This also helps reduce data redundancy.

- Data Reduction is the process of reducing data in size which can be performed in 2 ways: data size reduction based on rows and dimensionality reduction based on attributes. For particular techniques used in data mining, the same quantity of data is required to offer validity and accuracy of results.

2.3) Data Transformation is to convert information form one type to another applicable to particular models of data mining process. Transformation can take on the following methods.

- Normalization is to scale data to fall within specified ranges. This is to reduce data range because the same type of data but with too broad range may affect the analysis in particular data mining models. To give an example, data in a range of 0-1,000,000 can be scaled down to 0-20.

- Discretization is the process of transforming continuous numeric data into specific sub-ranges and treating each sub-range as a category in order to reduce data distribution, for example, score ranges of students, sub-ranges of customer's age: 20-25 years, 26-30 years, 31-35 years, respectively.

- 1 Of N Coding is a process of transforming analog data into digital data without identifying ascending or descending order such as in giving code to data.

- Generalization involves transforming too much detailed data into more concise terms, for example, customer's address may be abstracted as in Bangkok or in provinces.

3) Data Mining is the step of employing a range of techniques or operations such as database segmentation (clustering), data classification, link analysis (association), each of which offers different algorithms to be chosen from.

4) Interpretation/Evaluation or Analysis of Result and Knowledge Presentation. This final step requires data analysts to document or interpret results of data mining process as the new knowledge to be used as information to facilitate executives' decision making.

It can be said that data mining is a tool to assist in capturing hidden knowledge in massive amount of databases. The way in which the result is utilized depends on the analyst's decision to make use of it.

Data mining functionalities provide a range of applications. In general, there are two types of data mining tasks: predictive data mining and descriptive data mining. [9]

Predictive data mining attempts to do prediction of features or estimation of data values based on the past information, e.g. developing model upon the debt history of debtors in order to identify the characteristics of debtors with potential problematic debt. It is a technique used for predictive modeling.

Descriptive data mining aims to find a model that describes certain properties of available data. Mostly, data is categorized to describe the pattern of available data and serve to guide decision making.

2.5 Data Mining Technique

At present, data mining technique takes on several forms and it is not possible to identify which one is superior over another, however depending on functional objective. The following data mining techniques are commonly used. [9]

Classification or Predictive, This technique is used to classify data based on predetermined attributes. Classification technique partitions data into 2 segments: training data, and testing data. This process employs model construction (Learning) by learning from original data with predefined classes (Training data), and can handle both the categorical and numerical data for the model to learn behaviors of original data. The more the availability of training data, the better the accuracy of classification of the model. Classification or predictive techniques for data mining include decision trees and neural networks.

According to the techniques described above, the researcher adopted the Decision Tree approach in this study with an aim to find potential substances that yield the required hardness and softness of rubber, and to serve in modification of formula and the quantity of ingredients.

CHAPTER III
LITERATURE REVIEW

3.1 Decision Tree

Decision tree is a tree-like model as a Classification or Predictive model. Decision tree builds classification model using predetermined dataset, i.e. training set which is used to predict unclassified set of items. The basic principles [6] of creating a decision tree, Create a top to bottom (top-down) is from the root of the tree then branches to the leaves. Decision tree consists of different nodes with the topmost node called root node. Next to root node are branches representing decision choices. Leaf node represents the output of data prediction. Algorithms used include CHAID, CART, C4.5, for instances.

In the steps of building decision tree[16], algorithm C4.5, a widely known algorithm developed by Quinlan as an extension of his algorithm ID3 [18] provides a method of learning from the sample called training set, using classification to build decision tree. Training set is similar to the data in relational database presented in tables that contain rows of data or sample, and columns of data attribute. The attribute is of 2 types.

1) Goal attribute in data classification identifies the class to which a particular sample belongs. There is only 1 attribute in each dataset, and with only text data.

2) Predicting attribute identifies the attribute of each sample, and each attribute may include either text or numeric data.

3.2 Selection of attribute for data classification

To build the decision tree[16], a crucial problem needed to be considered is to make decision on what attribute should be selected to function as root node in each step of building decision tree and sub-trees. The criteria to facilitate attribute

selection are the calculation of information gain value as it indicates how well a given attribute can classify data. This can be done by choosing potential attributes from the dataset to function as root node. Any selected attribute that provides highest gain value can best classify the data, or the data in each leaf of the tree belongs to the same class or confounding with very little data from different class. Gain values for selecting attribute are described below.

3.2.1 Gain criterion

The method of decision tree[16] building using algorithm ID3 employs the gain criterion to make choice of attribute as root node or node of sub-tree by calculating for gain value of each attribute at classification time and choose the attribute with highest gain value as root node. The attribute chosen will be highly capable of classifying data with the need of smallest number of data to identify which group they belong to. Selection of attribute enables the separation of data with less confounding among different classes. Gain value is calculated by applying the knowledge from information theory implying that the expected value of information of data depends on the probability of information which is typically measured in bits, and can be written as:

Expected value of information = – log2 (probability of information) (3.1)

The use of gain value helps reduce the number of test time in classification of data, and also assure that the obtained decision tree is not too complex. The gain value can be computed from the following equation [6].

Expected gain value for data classification can be obtained by:

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s} \quad (3.2)$$

The value of entropy of attribute A with the value of attribute as (a1, a2, a3,..., av) is obtained by:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \ldots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj}) \quad (3.3)$$

Gain criterion used to select attribute A as node of decision tree is equal to the quantity of expected information to enable data classification subtracted by the quantity of expected information for data classification with attribute A to be tested for data classification, written as:

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A) \quad (3.4)$$

Given

- S is a dataset containing s record of information
- m is all of different classes of a dataset
- C_i is class i with i ranged between 1 to m
- S_i is the number of data belonging to S and in class C_i
- S_{ij} is the number of data belonging to S in class C_i by dividing information with potential value j of attribute A whereby j ranged between 1 to v
- *si/s* is probability that information is in class C_i

3.2.2 Gain ratio criterion

In algorithm ID3, gain criterion is mainly used for selecting attribute as root node of decision tree or sub-tree. However in algorithm C4.5, gain ratio criterion is added for making decision on attribute as root node since gain criterion becomes much biased towards the data containing attributes of many possible values such as dataset that comprises attribute of identification number with no duplicated values among sample. If data is classified using this particular attribute, then only 1 sample is obtained for 1 branch of tree Moreover, when entropy is calculated based on the classification with this attribute, the value obtained will be 0, hence giving this attribute highest gain value.

It can be observed that measurement of gain criterion will provide high value when particular attribute displays many possible values which cannot be used as node of the decision tree to accurately predict classes of new unseen data. Such bias therefore needs to be corrected by making adjustment on gain value for accuracy by using split information on attribute to calculate gain ratio criterion.

Given T represents training set when splitting sample with attribute A, the subsets of sample in each branch will be $\{t_1, t_2, \ldots, t_v\}$ at v sets as per possible values of attribute A, and split information is computed as:

$$\text{Information gain} = -\sum_{i=1}^{v} \frac{|t_i|}{|T|} \log_2\left(\frac{|t_i|}{T}\right) \qquad (3.5)$$

Split information value denotes the extent of data distribution when splitting sample T into v subsets according to possible values of attribute A, of which is highest when | ti| is 1 in every branch and become lower when | ti| increases. When the gain criterion is divided by split information value, the gain ration criterion is obtained which will help correct bias of gain criterion by reducing the gain ratio criterion of attribute with multiple possible values.

Gain ratio criterion = Gain criterion/split information

3.3 Pruning Decision Tree [16]

During the process of building decision tree, each branch may have undergone unusual growth because the training set contains noises as a result of error in data record or system itself or confounding outliers. Pruning is a technique to overcome such problem and help reduce the problem of over fitting. The problem may not rise with the decision tree structure in which classification goes well with the data set used to build the tree, however when dealing with new data the efficiency of classification reduces. Pruning relies on statistical values to decide to remove the least reliable branch so that the new tree works faster with increased capability of predicting new data more precisely. There are two types of pruning which are commonly used.

1) Pre-pruning

Pre-pruning is to remove branches or stop their growing during the process of building the decision tree by changing a removed node into a leaf and that such leaf represents class with the number of supported data or the probability of having data mostly belongs to that class.

During the construction of a decision tree, key statistical calculation or measurement is required which include such as $\chi 2$, information gain so that assessment can be made how to build the tree and braches. If the measured value of a node does not meet a given criteria, that node is considered not suitable to grow further. It is difficult to define the extent of criterion values, however with very high values the tree is too complex and too low values the tree is too small for functional application.

2) Post-pruning

Post-pruning is to remove branches of the completely built decision tree, by measuring the complexity of each node after pruning. The un-removed node at the bottom will be replaced by a leaf and represents class with greatest number of supported data.

For every non-leaf node, calculation is made for the value of expected error which represents any error occurs if the mode of sub-tree is removed. This calculation is performed by using the total value of error of each branch and weight proportionately to the branch. If the removal of particular mode leads to greater error, the node will be remained, and if the removal gives acceptable error, the node is then removed. After pruning, the accuracy of the tree is measured and the tree with lowest value of error will be selected.

Other than pruning by using measurement of error, other techniques are also available such as encoding based on Minimum Description Length (MDL).

The current research applies post-pruning because this method of pruning is more stable and effective than pre-pruning since it allows choosing pruning the useless node of the tree well-constructed by available dataset, and using various measurement methods to identify error of a node and consider if it is to be removed. Post-pruning takes on several forms such as **Reduced-error pruning** (REP),

Pessimistic error pruning (PEP), **Error-based pruning** (EBP). The researcher adopts the **Cost-complexity pruning** (CCP) approach.

Cost-complexity pruning [16]

Cost-complexity pruning (CCP) is a technique used in decision tree algorithm CART. The method consists of 2 steps.

1) Selecting set of sub-trees from the tree constructed by decision tree process T_{max} to obtain { T_0, T_1, T_2, … , T_L } where $T_0 = T_{max}$ and T_L is the root of decision tree.

2) Selecting the best tree T_i from the obtained set by evaluating the accuracy of the decision tree.

In the first step, the tree T_{i+1} derived from T_i by pruning that yields the lowest increase of resubstitution error rate when the tree T is pruned at node t, the increase error rate is therefore equals to R(t) - R(T_t) and the number of leafs reduced to $L(T_t)$ minus *1*.

$$\frac{R(t)-R(T_t)}{L(T_t)-1}=\alpha_T \tag{3.6}$$

The increase of resubstitution errors per the number of leafs being removed is called cost-complexity of the tree *T*', hence T_{i+1} derived from T_i by pruning the branch that yields lowest value of complexity, and in case the trees showing equal values of complexity, the one with less number of node will be selected.

It can be observed that the main class of sample shown by the leafs of sub-trees is negative. If the sub-trees are replaced by leafs of the negative class, the number of sample with classification error will be 1, Therefore, the classification error at this node, *R(t)*, is equal to *1/2514*. When considering the sub-trees with 4 leafs, it is seen that none of the sample displays inaccuracy classification. The cost-complexity can be computed as:

$$\alpha_T = \frac{(1/2514) - 0}{4 - 1} = 0.00013 \qquad (3.7)$$

Cost-complexity obtained is minimum, hence the tree *T1* derived from the tree under the construction steps *T0* by replacing these sub-trees with leafs will be selected and included in the set for further assessment of accuracy so as to yield decision tree that best classify new data. The decision tree undergone pruning with CCP method. Similarly, the second step is to choose the best tree by comparing the classification accuracy in order to find the error of each tree by using cross-validation method or the pruning data set for examination.

3.4 Cross-Validation

Cross-validation [15] is a method of predicting error of a model or resampling. The method starts from partitioning the dataset and examining some parts. Cross-validation result is often used as an option to define model. In using *K*-fold Cross Validation, the data is divided equally into *K* sets and calculated for error for *K* rounds. For each round the calculation is performed on a single set out of *K* sets which is selected as test set and the remaining *K-1* sets as training set, as shown in Figure 3.1.

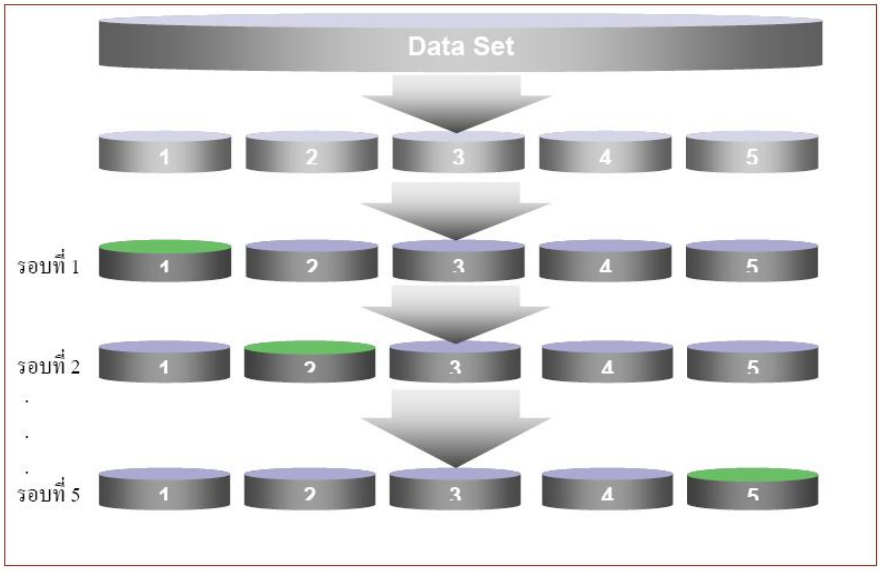

Figure 3.1 Partitioning dataset for calculation of error using cross-validation method.

3.5 Reliability Measurement

Once a model is constructed, it is subject to reliability test [15] by considering the statistical values obtained from the analysis and test as described next.

3.5.1 Confusion Matrix provides details of prediction results as in Figure 3.2.

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

Figure 3.2 Confusion Matrix displays analysis and test results

Where

a, d are the number of correct predictions

b, c are the number of incorrect predictions

3.5.2 Correctly Classified Instances represent correct predictions. Precision of calculation can be obtained from the following equation.

$$Correctly\ Classified\ Instances\ =\ \frac{(a+d)}{(a+b+c+d)} \tag{3.8}$$

3.5.3 True Positive rate (TP rate) is to indicate that the data as a result of prediction by the constructed Decision Tree Model is correctly identified, that is, the number of data correctly predicted in certain class/the total number of data in certain class, as calculated using the equation:

$$true\ positive\ (TP)\ rate = \frac{a}{(a+d)} \tag{3.9}$$

3.5.4 False Positive rate (FP rate) is similar to TP rate but differs in that False Positive refers to the data that does not belong to certain class but is incorrectly predicted by the model as a member of certain class, as calculated by the equation:

$$false\ positive\ (FP)\ rate \quad = \quad \frac{b}{(a+d)} \tag{3.10}$$

3.5.5 Root Mean Squared Error (RMSE) is an index used to measure the difference between the values actually observed and the values predicted by the model, as calculated by the equation:

$$RMSE = \sqrt{\frac{1}{n}\left\{\sum_{i=1}^{n}(o - o_{cal})^2\right\}} \tag{3.11}$$

Where

n is the total number of data

O is the values actually observed

O_{cal} is the value predicted by the model

3.6 Relevant Research

Data mining is currently applied to a wide range of fields other than business such as medicine, agriculture, education, and many more.

Kitti Chomchai and Montien Ratanasiriwongwut [11] adopted the Decision Tree technique in developing a system to support the choice decision to buy computers that satisfies the requirement and various aspects of functioning. The decision tree is employed to construct a model using algorithm SimpleCart and J48 and choose a model derived from algorithm SimpleCart to use in system development.

Suthinan Sangtong [12] applied data mining technique to the prediction of pig products by collecting data about pig raising. The prediction is divided into 2 parts, i.e. the part of classification and categorization of data based on quality grades of pig to which the decision tree was applied, and another part is the prediction of numeric data and various information on pig raising, using regression analysis, and the result was utilized in program development.

Chalermpol Takum and colleagues [13] used artificial neural networks in interpretation of cases expression towards Thai copyright Act which includes Thai word segmentation and interpretation of data. Artificial neural networks and decision tree technique were compared in order to develop a learning model with maximum precision. Results suggested that learning through the artificial neural networks provides better precision than the decision tree technique.

CHAPTER IV
RESEARCH METHODOLOGY

In the present research, the researcher applied data mining technique to the production of rubber shoe sole. The data employed includes the mixture of rubber and chemicals, softness-hardness, which will be analyzed to identify suitable values. To satisfy the objective, the research procedures, instruments, and methods are defined. This chapter describes in details on: 4.1) relevant information; 4.2) investigation of relevant articles; 4.3) formatting the data obtained for further analysis, setting out experimentation design and method; 4.4) experimentation; 4.5) accuracy test of model; and 4.6) report, analysis, and conclusion of results.

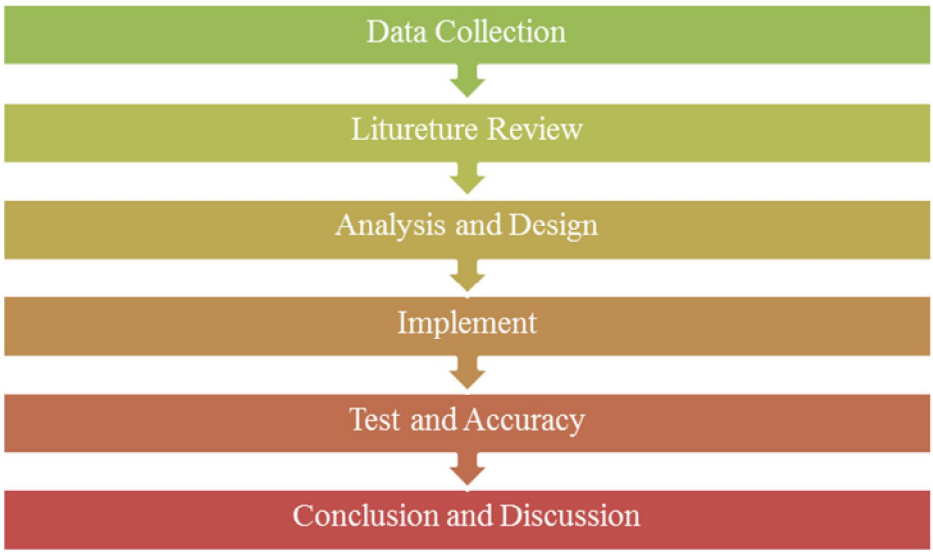

Figure 4.1 Step of work

4.1 Investigation of relevant information

The researcher carried out the data search to acquire rubber compound for making rubber shoe soles by investigating the following information.

4.1.1 Information on general conditions of the factory, production capacity, mixtures used in rubber processing. Since the case study is a shoe parts factory with specific production of shoe soles, a number of rubber formulation are thus available for this research.

4.1.2 Ingredients used for making shoe soles, types of rubber and chemicals, rubber compounding used and results of each formulation.

4.2 Investigation of relevant research

The focus is on research related to the designs, methods, and procedures of data mining, and research related to rubber formulation by using data mining technique

4.2.1 Relevant research and theories such as data mining, rubber formulation, and guides to data mining application.

4.2.2 Procedure and method of data mining, comparative test of the efficiency of techniques to be further selected for using in this research.

4.2.3 Instruments and methods used for data analysis. This research applied Weka program for data analysis since it offers many functions for statistical calculation and allows for modeling and displaying understandable results.

4.3 Data collection, selection, and preparation

Factory-based data, based on the data on general conditions of the factory, the factory produced 3 types of shoe soles, each of which uses different chemicals to prepare mixture. The researcher chooses to study rubber shoe sole because its production is of greater volume than other types and different formulations are more available. The following data are collected, selected and prepared.

Types of rubber

The rubbers used for making shoe soles are varied depending on their price and type, and properties. The choice of particular rubber type is thus important. Different types of rubber used are shown below.

Table 4.1 Types of rubber

Rubber	Price/Unit
SBR-1500	38.00
SBR-1502	38.00
SBR-1712	38.00
SBR-1716	38.00
Stipper	30.00
SKIM1	95.00
SKIM2	47.00
BR2	33.00
STR-5L	108.00
ADS	103.00
Reclaim	19.00
Latex	30.00

Chemicals

Chemicals are agents added into rubber to enhance its property as activator to provide rubber with required specification, such as the followings.

Table 4.2 Chemicals

Chemical	Price
C-75	1.16
SILICA	25.00
Coal tar	13.71
DEG	40.00
SP/C	45.00

Table 4.2 Chemicals (cont.)

Chemical	Price
ZINC OXIDE	37.50
WAX	57.00
STA^	35.00
MBTS	85.00

Ingredients

Ingredients are the data on rubber-chemical mixture used in manufacturing. The data contains the quantity of rubber and each substance used in the mixture.

Table 4.3 Ingredients

No.	Item	Weight	Price	Amount of money	PHR
1	BR2	38.00	33.00	1,254.00	95%
2	SKIM2	2.00	38.00	76.00	5%
Total rubber		**40.00**		**1,330.00**	**100%**
4	HISIL	18.00	25.00	450.00	45.0
5	SN.60	5.00	36.00	180.00	12.5
6	DEG	1.30	40.00	52.00	3.3
7	BHT	0.40	130.00	52.00	1.0
8	KALZIN	1.60	37.50	60.00	4.0
9	WAX	0.50	57.00	28.50	1.3
10	STA^	0.50	57.00	28.50	1.3
11	MBTS	0.40	85.00	34.00	1.0
Total rubber-chemicals			**67.70**	**2,215.00**	
Average price per kilogram				**32.72**	

Colors

Data on colors is obtained from mixing rubber and chemicals that produces colors. Colors are classified into different groups, i.e. black, raw, brown, chaff, and others.

Table 4.4 Colors

Brown		
No.	**Formula**	**Price/Kilogram**
1	นต.T220/1	20.64
2	นต.T400	27.78
3	นต.T403	22.76
4	นต.T420	43.09
5	นต.T433	47.28
6	ดิบT413	47.62
7	นต.T421	37.74

Softness-Hardness Values

Softness-hardness values are measured to identify the extent of softness and hardness of rubber shoe soles. The values differ among individual pairs of shoes despite using the same amount of ingredients due to a number of factors. Softness-hardness values are divided into ranges, i.e. 60-65, 65-70, 70-75, and 75-80.

4.4 Design and Implementation

This step involves constructing a model using the data collected. The current study adopts the Decision Tree approach to modeling for data classification. The algorithms used are C4.5 and CART. To create a model, the data is put into 2 groups: rubber and chemicals, which are separately tested and collectively tested. Based on the test results, the model is validated using K-fold Cross Validation.

The research applies Decision Tree to create a model so as to capture the suitable amount of rubber and chemicals used in the mixture for making rubber soles. The approach is widely used and the format of data is not much complicated. The structure of Decision Tree is familiar and easy to understand.

Implement step is creating a model as planned. Weka program is employed to build a model, which is the model of rubber compound to be used. Then, the next step is to test the accuracy and precision of the model.

4.5 Model validation

The model is examined for its accuracy using k-fold Cross Validation as a technique to assess the errors of predictive model. Normally, cross validation involves partitioning a sample of data into complementary subsets, and performing analysis on some subsets. Accordingly, the results are used as options for model identification. However, in the case of K - fold cross-validation, the original sample is randomly partitioned into K equal size subsamples. The cross-validation is repeated k times with each of the k subsamples used exactly once as the validation data for testing the model, and the remaining $k-1$ subsamples are used as training data. The model is test for reliability by using such as Confusion Matrix, Correctly Classified Instances, True Positive rate, False Positive rate, Root Mean Squared Error.

4.6 Report of results, analysis, and interpretation

This final step is to report the test results, analyze for the causes, interpretation of research results, and then make conclusion. The obtained information will become the prototype of mixture which can be further improved for better efficiency.

Experimentation instrument

Hardware: 1) Notebook Intel(R) Core(TM) I5 2.40 GHz RAM 8.00GB

Software: 1) Microsoft Windows7 2) Weka 3.6.9 3) Microsoft Excel 2010

CHAPTER V
RESULTS AND DISCUSSION

This chapter contain on the important issues as follow: 5.1 Data Study, 5.2 the result of Decision Tree to use algorithm J48, 5.3 the result of Decision Tree to use algorithm Simple CART, and 5.4 compare the Decision tree result from algorithm J48 and algorithm Simple CART by using Pair T-Test. And last is result discussion.

5.1 Data Study

Data of study is in this research manufacturing total sample is 56 sample Attributes is 33 Attributes divide to 32 Numeric Attributes and 1 Nominal Attributes, which Nominal Attributes is class of hardness to 4 class, namely, 75-80, 70-75, 60-65 and 65-70 shown in table 5.1.

Table 5.1 Characteristic of Data

Data Set	No. of Sample	No. of Attributes	No. of Nominal Attributes	No. of Numeric attributes	Missing values	No. of Classes
Rubber Formula	56	33	1	32	0	4

The analyze Manufacturing process of footwear to make different group of Hardness. The important factor to take into consideration is Rubber compounds because the properties of each type of rubber is not the same as the concentration of rubber. Quantity of chemicals used to produce the formula. Therefore, Separate data to 3 data sets are rubber data set, chemical data set and rubber and chemical data set. Attributes that shown in table 5.2

Table 5.2 Number of Attributes in each data set

Data	No. of Attributes
Rubber	13
Chemical	19
Rubber and Chemical	32

5.2 The result of J48 Tree

The 3 data set was created using the Model Tree J48 by measuring the accuracy of the system such as correctly classified, incorrectly classified, and root mean squared error that shown in chapter 2

5.2.1 Rubber Data Set

```
J48 pruned tree
------------------

STPIPER <= 0
|   RECLAIM <= 24
|   |   SKIM1 <= 0
|   |   |   SBR1712 <= 8
|   |   |   |   STR-5L <= 0
|   |   |   |   |   BR2 <= 85
|   |   |   |   |   |   SKIM2 <= 33: 70-75 (15.0/4.0)
|   |   |   |   |   |   SKIM2 > 33: 60-65 (2.0)
|   |   |   |   |   BR2 > 85: 60-65 (9.0/1.0)
|   |   |   |   STR-5L > 0: 60-65 (5.0)
|   |   |   SBR1712 > 8: 60-65 (2.0)
|   |   SKIM1 > 0: 60-65 (13.0/4.0)
|   RECLAIM > 24
|   |   SKIM2 <= 5
|   |   |   SBR1712 <= 0: 70-75 (2.0)
|   |   |   SBR1712 > 0: 75-80 (3.0/1.0)
|   |   SKIM2 > 5: 65-70 (3.0/1.0)
STPIPER > 0: 75-80 (2.0)

Number of Leaves  :     10

Size of the tree :      19
```

Figure 5.1 J48 tree model from rubber data set

Figure 5.1 show that J48 tree model have STPIPER is Root Node. The node minor is RECLAIM. The model have total Leaf Node are 10 Node and total branch have 19 Node.

5.2.2 Chemical data set

```
J48 pruned tree
------------------

C-45 <= 61
|   STA^ <= 1.4: 60-65 (27.0/1.0)
|   STA^ > 1.4
|   |   KALZIN <= 5
|   |   |   KALZIN <= 3: 65-70 (5.0/2.0)
|   |   |   KALZIN > 3: 70-75 (16.0/3.0)
|   |   KALZIN > 5
|   |   |   STA^ <= 4: 60-65 (2.0)
|   |   |   STA^ > 4: 70-75 (2.0/1.0)
C-45 > 61: 75-80 (4.0)

Number of Leaves  :     6

Size of the tree :      11
```

Figure 5.2 J48 tree model from chemical data set

Figure 5.2 show that J48 tree model have C-45 is Root Node, which have total Leaves Node are 6 Node and total branch have 11 Node.

5.2.3 Rubber and chemical data set

```
J48 pruned tree
------------------

C-45 <= 61
|   STA^ <= 1.4: 60-65 (27.0/1.0)
|   STA^ > 1.4
|   |   SKIM1 <= 0
|   |   |   KALZIN <= 3: 65-70 (5.0/2.0)
|   |   |   KALZIN > 3
|   |   |   |   KALZIN <= 5: 70-75 (13.0)
|   |   |   |   KALZIN > 5: 60-65 (3.0/1.0)
|   |   SKIM1 > 0: 65-70 (4.0)
C-45 > 61: 75-80 (4.0)

Number of Leaves  :     6

Size of the tree :      11
```

Figure 5.3 J48 tree model from rubber and chemical data set

Figure 5.3 show that J48 tree model. Root Node is C-45, which have total Leaves Node are 6 Node and total branch have 11 Node

5.2.4 Evaluation summary

After make decision tree from 3 samples, the important thing to consider is result of decision tree. That shows by accuracy value and error value which shows in table 5.3

Table 5.3 Summary of evaluation J48 model

Evaluation	Rubber Data	Chemical Data	Rubber & Chemical Data
Correctly Classified	80.36 %	87.5 %	92.86 %
Incorrectly Classified	19.64 %	12.5 %	7.14 %
Kappa Statistic	0.68	0.8	0.89
Mean Absolute Error	0.15	0.09	0.05
Root Mean Squared Error	0.27	0.22	0.16
Relative Absolute Error	45.52 %	29.69 %	16.97 %
Root Relative Squared Error	67.87 %	54.82 %	41.44 %

Table 5.3 show that the Rubber Data have correctly classified = 80.36 %, Incorrectly Classified = 19.64 % and root mean squared error = 0.27. The Chemical Data have correctly classified = 87.5 %, Incorrectly Classified = 12.5% and root mean squared error = 0.22. The Rubber and Chemical Data have correctly classified = 92.86 %, Incorrectly Classified = 7.14 % and root mean squared error = 0.16.

This table show result of Rubber and Chemical Data is the best value and process to produce the footwear use formula require a combination of rubber and chemicals therefore finding formula to produce footwear need to use both rubber and chemical.

All of results have good value because process to make Decision Tree does not validate data. The next step is validation data set.

5.2.5 Validation

Check the accuracy of model using by validation method is many methods. But the researchers have selected a K-Fold Cross Validation because the data used are little so they can't separate data for the Training Data Set and Testing Data Set which method Validation of K-Fold Cross Validation was already mentioned in the chapter 2.

```
J48 pruned tree
------------------

C-45 <= 61
|   STA^ <= 1.4: 60-65 (27.0/1.0)
|   STA^ > 1.4
|   |   SKIM1 <= 0
|   |   |   KALZIN <= 3: 65-70 (5.0/2.0)
|   |   |   KALZIN > 3
|   |   |   |   KALZIN <= 5: 70-75 (13.0)
|   |   |   |   KALZIN > 5: 60-65 (3.0/1.0)
|   |   SKIM1 > 0: 65-70 (4.0)
C-45 > 61: 75-80 (4.0)

Number of Leaves  :     6

Size of the tree :      11
```

Figure 5.4 Validation J48 tree model from rubber and chemical data set

Figure 5.4 show that validation J48 tree model has C-45 is Root Node, which have total Leaves Node are 6 Node and total branch have 11 Node.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          46               82.1429 %
Incorrectly Classified Instances        10               17.8571 %
Kappa statistic                          0.718
Mean absolute error                      0.0962
Root mean squared error                  0.2688
Relative absolute error                 30.0163 %
Root relative squared error             67.4534 %
Total Number of Instances               56
```

Figure 5.5 Stratified cross validation from rubber and chemical data set

Figure 5.5 show that the model have correctly classified = 82.14 %, Incorrectly Classified = 17.86% and root mean squared error = 0.27.

Validation of the results, the data show that the effect of a mixture of rubber and chemicals, which are correctly classified 82.14% on the Model and then analyzed the reliability of the model that we created by considering the statistical analysis of the Model.

Table 5.4 Confusion Matrix

Classified Class				Actual Class
A	B	C	D	
4	0	0	0	A = 75-80
1	**11**	3	1	B = 70-75
0	2	**26**	1	C = 60-65
0	2	0	**5**	D = 65-70

From Confusion Matrix, compare the predicted value with actual value by class such as TP Rate, FP Rate and ROC Area

where TP Rate is the number of data correctly predicted in certain class
FP Rate is the number of data incorrectly predicted in certain class
ROC Area is Area under Roc Curve

Table 5.5 Accuracy of model

Class	TP Rate	FP Rate	ROC Area
75-80	**1**	**0.019**	**0.99**
70-75	0.688	0.1	0.801
60-65	0.897	0.111	0.932
65-70	0.714	0.041	0.974
Weighted avg	0.821	0.093	0.904

Table 5.5 shows that TP Rate, FP rate and ROC Area of each Class in which the TP Rate Class 75-80 most valuable, followed by the class 60-65, 65-70 and 70-75, respectively. The FP Rate is sort of low class 75-80, 65-70, 70-75 and 60-65 in the ROC Area Class 75-80 the most valuable, followed by, 65-70, 60-65. , 70-75.

5.3 The result of SimpleCART

Algorithm CART algorithm is used for cutting branches (pruning) of the Decision Tree of the WEKA SimpleCART This technique is called "Cost-complexity pruning", which is described in Chapter 2.

5.3.1 Rubber data set

```
CART Decision Tree

RECLAIM < 38.0
|  SKIM2 < 14.0
|  |  SBR1502 < 19.0: 60-65(22.0/4.0)
|  |  SBR1502 >= 19.0: 70-75(2.0/2.0)
|  SKIM2 >= 14.0
|  |  SKIM2 < 19.0: 70-75(5.0/0.0)
|  |  SKIM2 >= 19.0
|  |  |  BR2 < 69.5
|  |  |  |  SKIM2 < 26.5: 60-65(4.0/0.0)
|  |  |  |  SKIM2 >= 26.5
|  |  |  |  |  SKIM2 < 34.5: 70-75(2.0/1.0)
|  |  |  |  |  SKIM2 >= 34.5: 60-65(2.0/0.0)
|  |  |  BR2 >= 69.5: 70-75(2.0/0.0)
RECLAIM >= 38.0
|  RECLAIM < 72.5: 75-80(3.0/1.0)
|  RECLAIM >= 72.5: 70-75(4.0/2.0)

Number of Leaf Nodes: 9

Size of the Tree: 17
```

Figure 5.6 SimpleCART Model fom rubber data set

Figure 5.6 show that root node is RECLAIM. Total nodes have 17 nodes. And Leaf Nodes have 9 nodes.

5.3.2 Chemical data set

```
CART Decision Tree

STA^ < 1.45: 60-65(26.0/1.0)
STA^ >= 1.45
|  C-45 < 114.05: 70-75(15.0/10.0)
|  C-45 >= 114.05: 75-80(4.0/0.0)

Number of Leaf Nodes: 3

Size of the Tree: 5
```

Figure 5.7 SimpleCART Model from chemical data set

Figure 5.7 show that root node is STA. total nodes have 5 nodes and Leaf Nodes have 3 nodes.

5.3.3 Rubber and chemical data set

```
CART Decision Tree

STA^ < 1.45: 60-65(26.0/1.0)
STA^ >= 1.45
|  C-45 < 114.05
|  |  SKIM1 < 7.5
|  |  |  KALZIN < 3.05: 65-70(3.0/2.0)
|  |  |  KALZIN >= 3.05: 70-75(14.0/2.0)
|  |  SKIM1 >= 7.5: 65-70(4.0/0.0)
|  C-45 >= 114.05: 75-80(4.0/0.0)

Number of Leaf Nodes: 5

Size of the Tree: 9
```

Figure 5.8 SimpleCART Model from rubber and chemical data set

Figure 5.8 show that root node is STA. total nodes are 9 nodes and Leaf Nodes are 5 nodes.

5.3.4 Evaluation summary

The important thing to consider is result of decision tree, which shows by accuracy value and error value that shows in table 5.6.

Table 5.6 Summary of evaluation SimpleCART model

Evaluation	Rubber Data	Chemical Data	Rubber & Chemical Data
Correctly Classified	82.14 %	80.36 %	91.07 %
Incorrectly Classified	17.86 %	19.64 %	8.93 %
Kappa Statistic	0.7	0.68	0.86
Mean Absolute Error	0.14	0.14	0.07
Root Mean Squared Error	0.26	0.26	0.19
Relative Absolute Error	42.95 %	43.71 %	23.04 %
Root Relative Squared Error	65.92 %	66.5 %	48.29 %

Table 5.6 show that the Rubber Data have correctly classified = 82.14 %, Incorrectly Classified = 17.86% and root mean squared error = 0.26. The Chemical Data have correctly classified = 80.36%, Incorrectly Classified = 19.64% and root mean squared error = 0.26. The Rubber and Chemical Data have correctly classified = 91.07 %, Incorrectly Classified = 8.93% and root mean squared error = 0.19.

The table show result of Rubber and Chemical Data is the best value like J48 model, and besides all of results have good value because process to make Decision Tree does not validate data. Then the test data set to valid.

5.3.5 Validation

The validation of rubber and chemicals data set.

```
CART Decision Tree

STA^ < 1.45: 60-65(26.0/1.0)
STA^ >= 1.45
|  C-45 < 114.05
|  |  SKIM1 < 7.5
|  |  |  KALZIN < 3.05: 65-70(3.0/2.0)
|  |  |  KALZIN >= 3.05: 70-75(14.0/2.0)
|  |  SKIM1 >= 7.5: 65-70(4.0/0.0)
|  C-45 >= 114.05: 75-80(4.0/0.0)

Number of Leaf Nodes: 5

Size of the Tree: 9
```

Figure 5.9 Validation SimpleCART model from rubber and chemical data set

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          45               80.3571 %
Incorrectly Classified Instances        11               19.6429 %
Kappa statistic                          0.6849
Mean absolute error                      0.1262
Root mean squared error                  0.2771
Relative absolute error                 39.3813 %
Root relative squared error             69.5365 %
Total Number of Instances               56
```

Figure 5.10 Stratified cross validation from rubber and chemical data set

Figure 5.9 and Figure 5.10 show STA is root node. Size of tree has 9 nodes and Leaf Node is 5 nodes. The model have correctly classified = 80.36 %, Incorrectly Classified = 19.64% and root mean squared error = 0.19.

Validation of the results, the data show that the effect of a mixture of rubber and chemicals, which are correctly classified 80.34% on the Model and then analyzed the reliability of the model that we created by considering the statistical analysis of the Model.

Table 5.7 Confusion Matrix

Classified Class				Actual Class
A	**B**	**C**	**D**	
4	0	0	0	A = 75-80
0	**14**	1	1	B = 70-75
0	3	**26**	0	C = 60-65
0	6	0	**1**	D = 65-70

From Confusion Matrix, compare the predicted value with actual value by class such as TP Rate, FP Rate, F-Measure and ROC Area

Table 5.8 Accuracy of model

Class	TP Rate	FP Rate	ROC Area
75-80	**1**	**0**	**1**
70-75	0.875	0.225	0.794
60-65	0.897	0.037	0.891
65-70	0.143	0.02	0.883
Weighted avg	0.804	0.086	0.87

Table 5.8 shows that TP Rate, FP rate, F-Measure, ROC Area of each class in which the TP Rate Class 75-80 most valuable, followed by 60-65, 70-75 and 65-70, respectively. FP Rate is minimal class 75-80, followed by 65-70, 60-65 and 70-75, respectively, the F-measure is the most inferior class 75-80 60-65, 70-75 and 65 -. ROC Area 70 the best is 75-80, followed by 60-65, 65-70 and 70-75.

5.4 Comparison the result between J48 and SimpleCART

Comparison to the second of these researchers are using the data to create the Model and then Validation using the 10-Fold Cross Validation and the calculation repeated 10 cycles using Paired T-Test in the WEKA Experimenter which the results can be shown in table 5.9.

Table 5.9 Comparison model

Algorithm	Percent Correct	RMSE	TP Rate	FP Rate	ROC Area
J48	**84.67**	**0.22**	**0.40**	0.02	**0.99**
SimpleCART	80.20	0.26	0.23	**0**	0.85

- J48 tree model accuracy was 84.67% above SimpleCART model accuracy is 80.20%.

- The RMSE J48 = 0.22 SimpleCART = 0.26 J48, which is less than the RMSE SimpleCART Model J48, the better, because if the RMSE low impact on the Model.

- The TP Rate FP Rate must be combined with the results obtained for J48 TP Rate = 0.40 SimpleCART = 0.23 and the FP Rate J48 = 0.02 SimpleCART = 0, so that from the Model J48 is more accurate.

- ROC Area is the value of the Plot ROC Curve during the TP Rate FP Rate on so much value that have accurate results and j48 = 0.99 SimpleCART = 0.85 J48 conclude that better SimpleCART.

5.5 Discussion

5.5.1 Data Extraction

From the objective to looking for the ingredients appropriately with soft rubber values due to the classification of data can analyzed the trend of data. This research is divided the ingredients into 3 samples which are rubber, chemical, and rubber and chemical. These patterns are used to test between rubber and chemical which are the most effected to soft rubber.

Table 5.10 Comparison result

Result	J48			SimpleCART		
	Rubber	Chemical	All	Rubber	Chemical	All
Percent correct	51.79	71.42	82.14	48.21	75	80.36
Percent incorrect	48.21	28.57	17.86	51.79	25	19.64
Root mean squared error	0.44	0.34	0.27	0.45	0.3	0.19

The Table 5.10 showed the comparison result of 2 models. The accuracy of chemical from both model are higher than the rubber, that shown the chemical effected to soft rubber more than rubber which are correspond to Table 2.4. The Table 5.2 was described the quantity of chemical added in the reinforcement type, there are in the range 10 – 100 phr. Then the chemical can be added to 100 percentage of quantity of rubber. The increment of chemical affected to the hardness of rubber is high. However, in the actually of footwear manufacture, the ingredients was used appropriately to get the hardness of rubber requirement. Thus, the chemical and rubber should be used together. The rubber is auxiliary to increase the hardness. From Table 5.10 illustrate the percentage of chemical and rubber was used together is high accuracy.

5.5.2 Data Modeling

The result from 2 models which are J48 and Simple CART Tree are difference. Thus, the selection model to match this objective must be considered the accuracy and suitability of model which the same data set.

J48 and SimpleCART are method to create the Decision Tree, there are described in Chapter 3. Although, the method of both are difference but the accuracy of both model are high equally. The J48 Model has the size of Tree is large. The number of variable is affected to the size of Tree, which is more variable the size of Tree is large. However, the accuracy of this model is higher than SimpleCART model. The size of Tree of SimpleCART model is smaller. This model is suitable to the large data. Though, the SimpleCART model is reduced the size of Tree, however the accuracy of this model remains high as well. The comparison result of 2 models which are J48 and SimpleCART model was showed in Table 5.9. Therefore, the selection of suitable model, in this research is used Paired T-Test to compare the effective of 2 models.

5.5.3 The Comparison of Data Modeling

From the results of Paired T-Test obtained above to the Model from J48 to the prediction, the better it may be due to SimpleCART as a way to cut the branches, so they focus on reducing the number of Node the size of the Tree is smaller. This results in very large systems. It can reduce the computation, but the accuracy of the system is or may be reduced slightly. Performance was not different. If we see from the data, the data is not very large size of the Tree is so small. Thus, the efficiency of the calculation is no different. It is necessary to measure at Accuracy of the Model J48 SimpleCART outperformed almost all the measures.

Therefore, it is concluded that the combination of a data set using the J48 is the most appropriate way to create the Model.

CHAPTER VI
CONCLUSION AND RECOMMENDATIONS

This chapter is divided into three topics to follow: 6.1 Summary of the methods and the results, 6.2 Recommendation and 6.3 Limitation of this research.

6.1 Conclusion

The research aims to design a model using by classification Decision Tree technique. The study showed that Decision Tree have percent of Correctly classified compared the Neural Network, Logistic Regression and Naïve Bayes.

In the experiments using Decision Tree technique to separate the data into three types due to the formulation of rubber compounds of tire to be first. It will lead to the other ingredients by percent compared to the first information is that two rubber is chemical. The final design is a combination of rubber and chemicals. How to compare between J48 and J48 SimpleCART found to have the accuracy 84.67% Accuracy 0.22 SimpleCART the way with 80.20% accuracy tolerances 12:26

Therefore, it can be concluded that to find a combination to suit the hardness of soft rubber soles. With J48 Decision Tree technique can be used to model due to the high accuracy and low tolerances. That show component in table 6.1

Table 6.1 Component

Hardness	Component
60-65	C-45 <= 61 / STA <= 1.4
	C-45 <= 61 / STA > 1.4 / KALZIN > 5
65-70	C-45 <= 61 / STA > 1.4 / SKIM1 > 0
	C-45 <= 61 / STA > 1.4 / KALZIN <= 3
70-75	C-45 <= 61 / STA > 1.4 / KALZIN > 3 & <= 5
75-80	C-45 > 61

6.2 Recommendation

The following recommendations were recommendations for improvement in further study is below:

1. Development Model using a variety of needs, such as price, color and so on.

2. Development Model is the program to continue to work on implementation.

3. Development Model can be applied to other parts of the following products.

6.3 Limitation of research

The results will be seen that the values obtained with high accuracy and low error, which sometimes caused by the sample used is minimal. When there is less to build and Measurement Model has a high threshold as well as competitions Model made of the data available. Model testing is so high, the opportunity to use the same data used to create the Field Model. The values obtained with high accuracy.

REFERENCES

1 อมร อาจินสมาจาร. (2553). *การศึกษาสมบัติเชิงกลของการผสมยาง HNBR/NBR เพื่อลดต้นทุนราคายางคอมเปาว์ด*. มหาวิทยาลัยธรรมศาสตร์.

2 กรมวิทยาศาสตร์บริการ. *ยาง*.http://lib3.dss.go.th/fulltext/board/ยาง1.pdf

3 คณะเทคโนโลยีสารสนเทศสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. (2547). Data Mining & Data Exploration Lab.

4 Roiger, R.J., &Geatz, M.W. (2003). *Data Mining A TUTORIAL-BASED PROMER*. United States of America: Pearson Education.

5 Larose, D.T. (2005). *Discovering Knowledge in Data*.NewJersey: John Wiley & Sons.

6 Han, J., &Kamber, M. (2001). *Data Mining Concepts and Techniques*. United States of America: Academic Press.

7 Cabena, P., Hadjnian, Stadler, Verhees&Zanasi. (1997). *Discovering data Mining: from concept to implementation.* Ney Jersey: Prentice-Hall.

8 วิไล วรรณเชรษญศิริ. (2551).มหาวิทยาลัยศิลปากร.

9 เรวดีศักดิ์ดุลยธรรม. (2553).*การใช้เทคนิคดาต้าไมน์นิงในการสร้างฐานความรู้เพื่อการทำนายสัมฤทธิ์ผลทางการเรียนของนักศึกษาวิทยาลัยราชพฤกษ์*.

10 วราชัย สิงห์จังหรีด. (2551). *การวิเคราะห์อาการเสียของฮาร์ดดิสก์โดยวิธีการเหมืองข้อมูล*. มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

11 กิตติ โฉมฉายและมณเฑียร รัตนศิริวงศ์วุฒิ. (2010). *การพัฒนาระบบสนับสนุนการตัดสินใจเลือกซื้อคอมพิวเตอร์โน้ตบุ๊คโดยใช้เทคนิค Decision Tree Classification*. มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

12 สุทธินันท์ สังข์ทอง. (2550) *ระบบวิเคราะห์และพยากรณ์ผลผลิตสุกรด้วยเทคนิคการทำเหมืองข้อมูล*. มหาวิทยาลัยเชียงใหม่.

13 เฉลิมพล ตาคำและคณะ. (2010). *การตีความสำนวนคดีตามพระราชบัญญัติลิขสิทธิ์ไทยโดยใช้โครงข่ายประสาทเทียม*. วิทยาลัยนอร์ทกรุงเทพ.

14 พงษ์ธร แซ่อุย (2548). *ยาง : ชนิด สมบัติและการใช้งาน*. ศูนย์เทคโนโลยีโลหะและวัสดุแห่งชาติ(เอ็ม เทค), ซีเอ็ดยูเคชั่น, กรุงเทพฯ.

15 ขจรศักดิ์ ศรีอ่อน.(2552) .*Prediction of Power Interruption Cause Using Data Mining Techniques for PEA Distribution Systems in Central 1 Area*. มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ.

16 Narupon Wongprachanukul. (2005). *A PROPER METHOD FOR DECISION TREE PRUNING IN SCIENTIFIC DATA MINING*. Suranaree University of Technology.

BIOGRAPHY

NAME	Mr. Parinya Wongchampa
DATE OF BIRTH	12 March 1984
PLACE OF BIRTH	Nakhon Pathom, Thailand
INSTITUTIONS ATTENDED	Prince of Songkla University, 2003-2007 Bachelor of Arts (Political Science) Mahidol University, 2008-2012 Master of Science (Technology of Information System Management)
HOME ADDRESS	116 Moo 3 Banlaung, Dontoom, Nakhon Pathom 73150 Email: benz-zaa@hotmail.com