APPLICATION OF DATA MINING USING DECISION TREE METHOD FOR MONITORING THE DISTRIBUTION OF MOTORCYCLE SALES IN PD. WIJAYA ABADI BANDUNG

JAKA ADI SWARA

Jurusan Teknik Informatika – Universitas Komputer Indonesia

Jl. Dipatiukur No. 112-114 Bandung 40132

jaka_4s@yahoo.co.id

Pembimbing : Edi Mulyana, M.T.

ABSCTRACT

PD. Wijaya Abadi Bandung is the company engaged in buying, selling and also distributes Honda motorcycles branded to individual consumers and companies. As the continued passing of the transaction process of distribution, data will continue to form until the piles so that difficult company to analyze data in order to improve distribution of motorcycle strategies. From these data the company wants any further data processing to find information / new knowledge useful. Desired information is the classification of the company is the relationship of a target company attributes its distribution with the motor type and color codes that exist in motorcycle sales distribution data.

To get such information more quickly and efficiently we should bear in an information technology support, in this case is data mining. Data mining is a method of information search (knowledge) which is contained in the data is very large. Deicision Tree is one of classification and prediction methods are very powerful and famous in the application of data mining. The selection method is based on the information you want to obtain the information that is classified. The advantage of this method is effective in analyzing a large number of attributes of the data available and easily understood by end users.

The output of this application is the information that is classified in the form of Decision Tree. This application can test the value of the attribute to determine the distribution pattern of the existing rules so that the company has a stock of knowledge in distributing motorcycles to the company's distribution target.

***Keywords** : Knowledge, Data Mining, Attributes, Classification, Decision Tree*

1. INTRODUCTION

PD. Wijaya Abadi Bandung is the company engaged in buying and selling branded Honda motorcycle. In the normal operaisonalnya, PD. Wijaya Abadi also distributes consumer motorcycle against individuals and companies. Each distributes motorcycles, the company usually records the transaction data distribution motorcycle into a database.

In one day, this company could be a lot of transactions, especially on the distribution of motorcycle sales. Because the numbers very much, the data difficult to analyze. From the data that is very much the company wants any further data processing to find information / new knowledge that is useful as an aide in making decisions, in this case is distributing motorcycles to the company's distribution target. Information needed is to classify the relationship of a target company attributes its distribution by type of motor and color code on motorcycle sales distribution data stored in a database Wijaya Abadi, so the company has a stock of knowledge and rules of distribution.

To get such information more quickly and efficiently we should bear in an information technology support, in this case is data mining. Target data to be in the distribution of data-mining is a motorcycle sales in PD. Wijaya Abadi.

Data mining is a method of information search (knowledge) which is contained in the data is very large. In data mining there are several methods of prediction, classification, clustering, and association. Deicision Tree is one of classification and prediction methods are very powerful and famous in the application of data mining. Basically Decision Tree turns data into decision tree (decision tree) and decision rules (rule). The advantage of this method is effective in analyzing a large number of attributes of the data available and easily understood by end users.

The purpose of this report is to build Application to Data Mining Using Decision Tree Method For Monitoring Motorcycle Sales Distribution in PD. Wijaya Abadi Bandung. With this application, it is expected to:

1. Provides information on the classification of motorcycle sales distribution using data mining in the form of a decision tree (decision tree) in PD. Wijaya Abadi Bandung.
2. Giving stock of knowledge and rules in making decisions distribute motorcycles in PD.Wijaya Abadi to companies target distribution.
3. Produce the pattern of linkages between attributes tipe_motor, kode_warna against ness is enforced the rules used in the distribution of motorcycles.

2. MODEL, ANALYSIS, DESIGN AND IMPLEMENTATION

The research was conducted using descriptive analysis method is a method that aims to get a clear picture about the things that need through the following stages:

a. **Data Collection**, consisting of: literature studies, field studies, and interviews.
b. **Software Development**, consisting of: Systems Engineering, Systems Analysis, System Design, Coding System, System Testing, Maintenance Systems, and feedback.

2.1 Data Analysis

Database of PD.Wijaya Abadi divided into several tables, outlines the tables are color table, table TipeMotor, tables Companies and as the target data in the application of data mining techniques is the target company's distribution stored in the table DistribusiMotor. Each variable / attribute decision making represents a particular dimension, in this case the dimensions involved are DsitribusiMotor. Motor Distribution table structure can be given in Table 2.1.

Table 2. 1 Structure of DistribusiMotor Table

Tabel DistribusiMotor

TIPE_MOTOR
KODE_WARNA
TGL_DISTRIBUSI
NO_DISTRIBUSI
TGL_PENJUALAN
NAMA_PERUSAHAAN

The data are incomplete and inconsistent and generally occurs in almost every database, incomplete data can be caused by a variety of reasons, such as the attributes with the wrong data. Similarly, the PD database. Wijaya Abadi, there are some attributes that are not really necessary so that the process of Data Preprocessing is necessary so that the database in accordance with the provisions required by the system.

Data preprocessing is important in the process of data mining, things that are included in it are:

a. ***Data Selection***

Before entering into the Data Preprocessing, which should be done early is the selection of data (data selection). On Eternal PD.Wijaya database, the data will be processed for the data-mining DistribusiMotor. This DistribusiMotor data will become Case Data in the operational processes of data mining. From existing data, the column is taken as attribute / decision variable is a column ness is enforced, while the column is taken as a determinant variable in the formation of a decision tree is a column:

1. **tipe_motor**
2. **kode_warna**

Selection of these variables with the consideration that the total value of the variables are not many and not unique so expect a motorcycle fit in a single classification variable value is pretty much.

b. ***Data Preprocessing / Data Cleaning***

Data cleaning is applied to add the content attribute is missing or empty, and change data that is not consistent. Stages in the data cleaning DistribusiMotor applied to the table that will be used as input for data mining process itself is as follows:

1. Ignore the tuple, performed if the value of a tuple is missing or does not exist. This method is not effective when there are many attributes with empty tuples.
2. Adding content to an empty attributes manually, but this approach is extremely time consuming and not effective when applied to very large data.

c. ***Data Transformation***

In this process, data is transformed into a form suitable for data mining process.

d. ***Data Reduction***

Data reduction is done by eliminating those attributes that are not needed so that the size of the database to be small and include only the attributes needed in the process of data mining. This is because the process of data mining will be more efficient on smaller data. From the data reduction process is then made the table with the attributes that are ready for the process of classification, in this case the CASE table as in Table 2.2.

Tabel 2. 2 Structure of KASUS Table

Tabel KASUS

TIPE_MOTOR
KODE_WARNA
NAMA_PERUSAHAAN

From the process of data preprocessing on the resulting table is ready for the process of classification, with attributes and attribute values listed in Table 2.3.

Table 2. 3 Results Process of Data Preprocessing

TIPE_MOTOR	KODE_WARNA	NAMA_PERUSAHAAN
GL160D2	AH	GELORA FAJAR PERKASA PT
GL160CW2	AR	GEMILANG MAKMUR ABADI CV
GL200R1	BB	
NC110CW1	BE	
NC110CW2A/T	BG	
NC11A2CA/T	BK	
NC11BICA/T	BL	
NC11B2C	BP	
NC11B3CA/T	BR	
NC11B3DA/T	BS	
NF11A1CM/T	BZ	
NF11A1CAM/T	GR	
NF11B1CA	GS	
NF11BICM/T	HA	
NF11BIDM/T	HB	
NF11B1D1M/T	HJ	
NF11B1C1M/T	HS	
NF125TR1	HZ	
NF125TR2	LS	
NF125TRF1	MH	
NF125TRF2	MS	
NF125TD1	NK	
NF125TD2	OR	
CS12A1RR	OU	
	PH	
	PJ	
	PM	
	PP	
	PK	
	PS	
	RD	
	RS	
	SB	
	SH	
	SV	
	UO	
	VS	
	WG	
	WH	
	WP	

Tipe_motor attribute usually contains the value in the types of Honda motor. The value of the attribute tipe_motor above can be explained in Table 2.4.

Tabel 2. 4 Tipe Motor Attribute Description

TIPE_MOTOR	Nama Motor
GL160D2	MEGA PRO D CW
GL160CW2	MEGA PRO SPOKE
GL200R1	TIGER CW
NC110CW1	VARIO CW

NC110CW2A/T	NEW VARIO CW
NC11A2CA/T	VARIO TECHNO
NC11BICA/T	NEW BEAT
NC11B2C	NEW BEAT CW
NC11B3CA/T	NEW BEAT SPOKE
NC11B3DA/T	BEAT
NF11A1CM/T	BLADE REPSOL
NF11A1CAM/T	BLADE
NF11B1CA	ABSOLUTE REVO
NF11BICM/T	ABSOLUTE REVO DELUXE
NF11BIDM/T	NEW ABSOLUTE REVO
NF11B1D1M/T	ABSOLUTE REVO CW
NF11B1C1M/T	ABSOLUTE REVO SPOKE
NF125TR1	SUPRA X 125 D
NF125TR2	SUPRA X 125 DD
NF125TRF1	SUPRA X 125 CW
NF125TRF2	SUPRA X 125R CW
NF125TD1	SUPRA X 125 INJ
NF125TD2	SUPRA X 125R INJ
CS12A1RR	CITY SPORT ONE

As for this kode_warna attribute contains a combination of two colors. The value of the attribute kode_warna in Table 2.3 can be explained in Table 2.5.

Tabel 2. 5 Kode Warna Attribute Description

KODE_WARNA	Nama Warna
AH	ABU HITAM
AR	ABU RED
BB	BLACK
BE	BLACK E
BG	BLACK GREY
BK	BLACK K
BL	BLUE
BP	BIRU PUTIH
BR	BLACK RED
BS	BLACK SILVER
BZ	BLACK Z
GR	GREY RED
GS	GREY SILVER
HA	HITAM ABU
HB	HITAM BIRU
HJ	HITAM MERAH
HS	HITAM SILVER
HZ	HITAM Z
LS	LIGHT SILVER
MH	MERAH HITAM
MS	MERAH SILVER
NK	NILA KUNING
OR	ORANGE
OU	ORANGE UNGU
PH	PUTIH HITAM
PJ	PUTIH JINGGA
PM	PUTIH MERAH
PP	PUTIH PINK
PK	PUTIH KUNING
PS	PUTIH SILVER
RD	RED
RS	RED SILVER
SB	SILVER BIRU
SH	SILVER HITAM
SV	SILVER VIOLET
UO	U ORANGE
VS	VIOLET SILVER
WG	WHITE GREEN
WH	WHITE
WP	WHITE PINK

2.2 Method of *Decision Tree* Analysis

Preprocessing of the data that has been done then produced a table that is ready for the process of classification, as in Table 2.6.

Tabel 2. 6 KASUS Table

No	tipe_motor	kode_warna	nama_perusahaan
1	GL200R1	BE	GELORA FAJAR PERKASA PT
2	NC110CW1	WG	GELORA FAJAR PERKASA PT
3	NC11BICA/T	NK	GELORA FAJAR PERKASA PT
4	NC11BICA/T	NK	GELORA FAJAR PERKASA PT
5	NC11BICA/T	BK	GELORA FAJAR PERKASA PT
6	NF125TR1	HJ	GELORA FAJAR PERKASA PT
7	NF125TR1	BG	GELORA FAJAR PERKASA PT
8	NF125TR1	BG	GELORA FAJAR PERKASA PT
9	NF125TR1	HJ	GELORA FAJAR PERKASA PT
10	NF11BICM/T	RD	GELORA FAJAR PERKASA PT
11	NF11BICM/T	BK	GELORA FAJAR PERKASA PT
12	NF11BICM/T	BK	GELORA FAJAR PERKASA PT
13	NF11BIDM/T	BB	GELORA FAJAR PERKASA PT
14	NF11BIDM/T	BB	GELORA FAJAR PERKASA PT
15	NF11BIDM/T	BS	GELORA FAJAR PERKASA PT
16	NF11BIDM/T	BS	GELORA FAJAR PERKASA PT
17	NF11BIDM/T	BR	GELORA FAJAR PERKASA PT
18	NF11BIDM/T	BR	GELORA FAJAR PERKASA PT
19	NF11BICM/T	RD	GELORA FAJAR PERKASA PT
20	NF11BICM/T	RD	GEMILANG MAKMUR ABADI CV
21	NC110CW1	WP	GEMILANG MAKMUR ABADI CV
22	NC110CW1	BK	GEMILANG MAKMUR ABADI CV
23	NC11BICA/T	NK	GEMILANG MAKMUR ABADI CV
24	NF11BICM/T	BK	GEMILANG MAKMUR ABADI CV
25	NF11BIDM/T	BB	GEMILANG MAKMUR ABADI CV
26	NF11BIDM/T	BR	GEMILANG MAKMUR ABADI CV

Table CASE is a collection of data to be processed, the above data is 26 data samples taken at random from the table DistribusiMotor. Attribute goals / decisions that ness is *nama_perusahaan*, which has two different content that is GELORA FAJAR PERKASA PT and GEMILANG MAKMUR ABADI CV that describes a company that became the target distribution. We assume that the class C_1= GELORA FAJAR PERKASA PT and C_2= GEMILANG MAKMUR ABADI CV, there are 19 sample data with the class of C_1 and 7 samples data with the class C_2.

Classification process begins by checking the contents of the attribute data class (ness is enforced) because the contents of different attributes are then used to obtain information entropy function highest (highest information gain) was calculated as follows:

$$I(s_1,s_2) = I(19,7) = -\frac{19}{26}log_2\frac{19}{26} - \frac{7}{26}log_2\frac{7}{26} = 0.84$$

1. Atribut tipe_motor

a. GL200R1

$$I(s_{11},s_{12}) = I(1,0) = -\frac{1}{1}log_2\frac{1}{1} - \frac{0}{1}log_2\frac{0}{1} = 0$$

b. NC110CW1

$$I(s_{21},s_{22}) = I(1,2) = -\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3} = 0.91$$

c. NC11BICA/T

$$I(s_{31},s_{32}) = I(3,1) = -\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4} = 0.81$$

d. NF125TR1

$$I(s_{41},s_{42}) = I(4,0) = -\frac{4}{4}log_2\frac{4}{4} - \frac{0}{4}log_2\frac{0}{4} = 0$$

e. NF11BICM/T

$$I(s_{51},s_{52}) = I(4,2) = -\frac{4}{6}log_2\frac{4}{6} - \frac{2}{6}log_2\frac{2}{6} = 0.92$$

f. NF11BIDM/T

$$I(s_{61},s_{62}) = I(6,2) = -\frac{6}{8}log_2\frac{6}{8} - \frac{2}{8}log_2\frac{2}{8} = 0.81$$

$$E(\text{tipe_motor}) = \frac{1}{26}I(s_{11},s_{12}) + \frac{3}{26}I(s_{21},s_{22}) + \frac{4}{26}I(s_{31},s_{32}) + \frac{4}{26}I(s_{41},s_{42}) + \frac{6}{26}I(s_{51},s_{52})\frac{8}{26}I(s_{61},s_{62})$$

$$= \frac{1}{26}0 + \frac{3}{26}0.91 + \frac{4}{26}0.81 + \frac{4}{26}0 + \frac{6}{26}0.92\frac{8}{26}0.81$$

$$= 0.691$$

$$Gain(\text{tipe_motor}) = I(s_1,s_2) - E(\text{tipe_motor})$$

$$= 0.84 - 0.691 = 0.15$$

Entropy value for the attribute tipe_motor generated is 0.691, which means that the level of diversity for the value of atirbut tipe_motor is 0.691. As for the value of gain or value effective to classify the attributes tipe_motor 0.15.

2. Atribut kode_warna

a. BE

$$I(s_{11},s_{12}) = I(1,0) = 0$$

b. WG

$$I(s_{21},s_{22}) = I(1,0) = 0$$

c. NK

$$I(s_{31},s_{32}) = I(2,1) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.92$$

d. BK

$$I(s_{41},s_{42}) = I(3,2) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.97$$

e. BS

$$I(s_{51},s_{52}) = I(2,0) = 0$$

f. BR

$$I(s_{61},s_{62}) = I(2,1) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.92$$

g. HJ

$$I(s_{71},s_{72}) = I(2,0) = 0$$

h. BG

$$I(s_{81},s_{82}) = I(2,0) = 0$$

i. RD

$$I(s_{91},s_{92}) = I(2,1) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.92$$

j. BB

$$I(s_{101},s_{102}) = I(2,1) = -\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3} = 0.92$$

k. WP

$$I(s_{111},s_{112}) = I(0,1) = 0$$

$$E(\text{kode_wama})$$

$$= 0 + 0 + \frac{3}{26}0.92 + \frac{5}{26}0.97 + 0 + \frac{3}{26}0.92 + 0 + 0 + \frac{3}{26}0.92 + \frac{3}{26}0.92 + 0$$

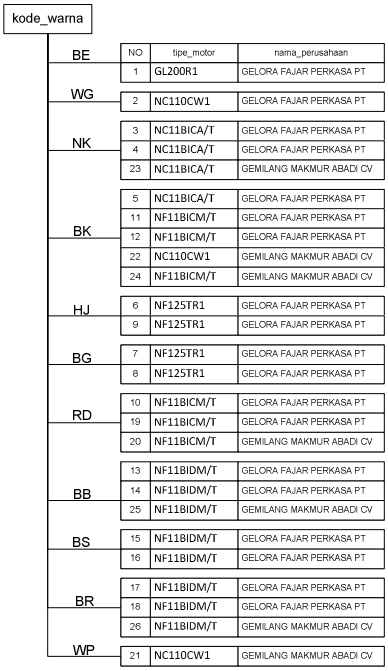
$$= 0.61$$

$$G(\text{kode_wama}) = (s_1,s_2) - E(\text{kode_wama})$$

$$= 0.84 - 0.61 = 0.23$$

Entropy value for the attribute tipe_motor produced is 0.61, which means that the level of diversity for the value of atirbut tipe_motor is 0.691. As for the value of gain or value effective to classify the attributes tipe_motor at 0.23.

From the above calculation it is found that the attribute kode_warna has the highest information (gain) is 0.23 compared with tipe_motor attribute, attribute kode_warna thus selected as the attribute to be tested. Where a node created with the label kode_warna and growing branch of the node based on the content attribute kode_warna and data partitioned based on the content attribute. Formed a decision tree can be shown in Figure 2.1.



	NO	tipe_motor	nama_perusahaan
BE	1	GL200R1	GELORA FAJAR PERKASA PT
WG	2	NC110CW1	GELORA FAJAR PERKASA PT
NK	3	NC11BICA/T	GELORA FAJAR PERKASA PT
	4	NC11BICA/T	GELORA FAJAR PERKASA PT
	23	NC11BICA/T	GEMILANG MAKMUR ABADI CV
BK	5	NC11BICA/T	GELORA FAJAR PERKASA PT
	11	NF11BICM/T	GELORA FAJAR PERKASA PT
	12	NF11BICM/T	GELORA FAJAR PERKASA PT
	22	NC110CW1	GEMILANG MAKMUR ABADI CV
	24	NF11BICM/T	GEMILANG MAKMUR ABADI CV
HJ	6	NF125TR1	GELORA FAJAR PERKASA PT
	9	NF125TR1	GELORA FAJAR PERKASA PT
BG	7	NF125TR1	GELORA FAJAR PERKASA PT
	8	NF125TR1	GELORA FAJAR PERKASA PT
RD	10	NF11BICM/T	GELORA FAJAR PERKASA PT
	19	NF11BICM/T	GELORA FAJAR PERKASA PT
	20	NF11BICM/T	GEMILANG MAKMUR ABADI CV
BB	13	NF11BIDM/T	GELORA FAJAR PERKASA PT
	14	NF11BIDM/T	GELORA FAJAR PERKASA PT
	25	NF11BIDM/T	GEMILANG MAKMUR ABADI CV
BS	15	NF11BIDM/T	GELORA FAJAR PERKASA PT
	16	NF11BIDM/T	GELORA FAJAR PERKASA PT
BR	17	NF11BIDM/T	GELORA FAJAR PERKASA PT
	18	NF11BIDM/T	GELORA FAJAR PERKASA PT
	26	NF11BIDM/T	GEMILANG MAKMUR ABADI CV
WP	21	NC110CW1	GEMILANG MAKMUR ABADI CV

Figure 2. 1 Phase -1 Classification Process

To fill attribute kode_warna BE, WG, HJ, BG, BS, and WP have the same sample data class, then it can be taken as a decision as depicted in Figure 2.2.

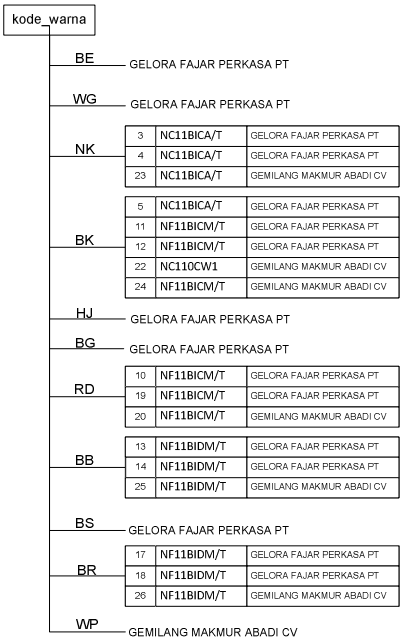


Figure 2. 2 Phase-2 Classification Process

From the above classification process can be obtained information that the content attribute kode_warna BE, WG, HJ, BG, and BS is usually distributed to the company of GELORA FAJAR PERKASA PT and **kode_warna** WP is usually didstribusikan to GEMILANG MAKMUR ABADI CV.

Therefore, determining which attributes are used only two attributes, namely kode_warna and tipe_motor the classification process stops at the stage-2, because it has no other attributes to be classified. So to kode_warna NK, BK, RD, BB, and BR unclassified.

2.3 DFD (*Data Flow Diagram*)

Modeling in this application using Data Flow Diagram using Power Designer tool.

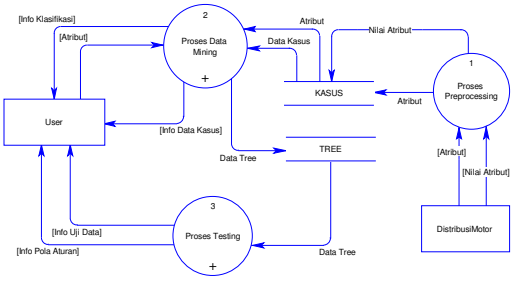


Figure 2.3 DFD Level 1

2.4 Structure Design Program

Design of Data Mining Applications program structure is as follows:

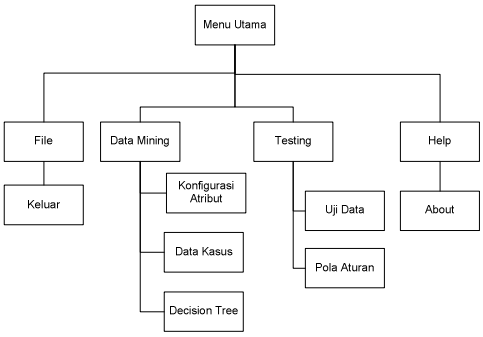
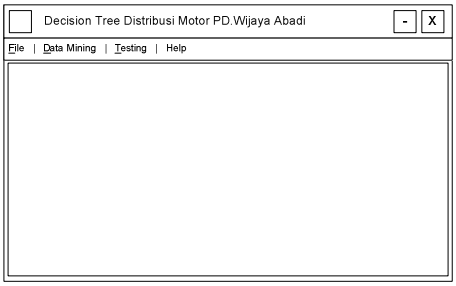


Figure 2.4 Structure of Menu

2.5 Interface Design



2.6 Implementation

Implementation stage is the stage of software creation, continuation phase of system design activities. This stage is the stage where the system is ready to operate.

Hardware equipment (hardware) and software (software) used in building this application are:

Table 2.7. Hardware Used

Hardware	Specification
Processor	Intel Core2Duo 1,5 GHz
RAM	1 GB
Harddisk	*Free memory 4GB*
Monitor	Resolusi 1024 x 768

Table 2.8 Software Used

No	Software	Description
1	Sistem Operasi	Windows XP
2	Bahasa Pemrograman	Pascal
3	*DBMS*	Interbase
5	*Code editor*	Borland Delphi 7
6	DFD modeler	Power Designer 6

3. RESULTS AND DISCUSSION

Tests are used to test this application is the method pengujina black box. Black box testing focuses on the functional requirements of software.

Based on the results of alpha testing has been done can be concluded as follows:

1. Still allow for errors in syntax as the maximum number of processes not yet invented.
2. Functionally, the system has to produce the expected output.

From the results of beta testing has been done is to test the calculation of choices from the questionnaire response categories that have been distributed in the field, it can be concluded that the Application Data Mining for Monitoring Distribution This motor allows the user to obtain information on the classification of a target company pendistribusianya with the attributes in the data distribution motorcycle sales more quickly and efficiently.

4. CONCLUSIONS AND SUGGESTIONS

Based on the results of the testing that has been done, it can be concluded as follows :

1. Data Mining Applications for monitoring the distribution of motorcycle sales in Bandung Eternal PD.Wijaya this can result in information classification attributes relationship to the data distribution to the company's motorcycle sales of its distribution targets in the form of a decision tree (decision tree) so as to provide stock of knowledge and rules in the decision distribution of motorcycles.
2. Decision Tree Using this method, the resulting information is classified and in accordance with the needs, and easily understood by users of the application in PD. Wijaya Eternal Bandung.

The system built is still a shortage, both in terms of functionality and data held. In making application to Data Mining Decision Tree method for monitoring Sepda Motor Sales Distribution In PD. Wijaya Abadi Bandung is still much that can be developed, such as:

1. Applications that have been built, its database using MySQL, it should be handled if a large amount of data that can be developed using the Oracle database.
2. This application should be equipped with features ¬ backup data, in order to cope with the event data is corrupt or lost data.

5. REFERENCES

[1] Feri Sulianta, Dominikus Juju, (2010), Data Mining Meramalkan Bisnis Perusahaan, Jakarta : Elex Media Komputindo.

[2] Hakim, Lukmanul, (2008), Membongkar Trik Rahasia Para Master PHP, Yogyakarta : Lokomedia.

[3] Hakim, Lukmanul, (2009), Trik Rahasia Master PHP Terbongkar Lagi, Yogyakarta : Lokomedia.

[4] Kadir, Abdul, (2005), Pemrograman Database Dengan Delphi 7, Yogyakarta : ANDI.

[5] Kadir, Abdul, (2003), Pengenalan Sistem Informasi, Yogyakarta : ANDI.

[6] Kusrini, Emha Taufiq Luthfi, (2009), Algoritma Data Mining,
Yogyakarta : ANDI.

Pressman, Roger, (2001), *Software Engineering : A Practitioner's Approach*, The McGraw-Hill Companies, Inc.