On-Time Flight Departure Prediction System Using Naive Bayes Classification Method (Case Study: XYZ Airline)

Andi Nugroho[1],Rizki Ali Fahmi[2]
[*]Informatics Engineering, Computer Science Faculty, Mercu Buana University

Abstract *- On Time Performance (OTP) is an important aspect for flight service user and provider. OTP is one of factors that affect positive or negative assessment of flight service. But sometimes, there are some obstacles happened that require the airlines to experience delay. Lack of information about delay prediction causes the airlines could not prepare the solution to avoid the delay problem. To overcome the problem, it requires a departure on time prediction system. In this research, the writer tries to apply Naive Baye Classification to create on time prediction system that can be used by the airlines to prepare more for the possibilities that can be happened in the future.*

Keywords - on time performance, Delay, Naive Bayes

I. INTRODUCTION

A. BACKGROUND

Soekarno-Hatta International Airport (IATA: CGK), is the main airport in Jakarta, Indonesia and located in Cengkareng, Tangerang. The airport start operations in 1985, replacing Kemayoran Airport. Soekarno-Hatta Airport is managed by PT. AngkasaPura II and serves about 45 airlines both from outside and within the country. In 2011, Soekarno-Hatta Airport served the 4th largest passenger in Asia after airports in Beijing, Tokyo and Hongkong, and ranked 12th in the world.The development of sensor nodes by considering multiple objectives and existence of fixed obstacles is an important optimization problem(Syarif, Abouaissa, Idoumghar, Sari, & Pascal Lorenz, 2014).The largest percentage of 97.69% of internet used to send and receive email, while the lowest is hotel promotion followed by VoIP with each percentage of 0.14% and 13.54%. (Bahaweres, Alaydrus, & Wahab, 2012)

The busyness of air traffic at Soekarno-Hatta Airport could cause the possibility of flight delays. Therefore need a system to predict on-time departure of flight departure so that the airlines can prepare themselves to handle the problems. With the preparation, no delay expected or can reduce the time of delay.

In this research will be discussed about the prediction system of punctuality of flight departure using naive bayes classification method. The classification method is a suitable method used in prediction systems, and naive bayes are classifications that have high speed and accuracy.

B. FORMULATION OF PROBLEM

1. How to predict the punctuality of flight departure using Naive Bayes classification?
2. How to process predicted data to be displayed in dashboard on web application?

C. LIMITS OF RESEARCH

1. Data used is dummy data xyz airlines.
2. Train data is flight data from Soekarno-Hatta airport during January 2016.
3. Test data is flight data from Soekarno-Hatta Airport on January 2-8, 2017.
4. The method used is Naive Bayes Classification.
5. Results of the classification will be displayed in a web-based application.

D. OBJECTIVE AND BENEFITS

The objectives to be achieved by researchers are:

1. Predict the punctuality of flight departure using the Naive Bayes classification method.
2. Processing the result data to be displayed in the dashboard on web-based applications.

Researchers hope that this research can provide some benefits that are:

1. Provide information to the airlines about the prediction of the punctuality of flight departure.
2. Become an evaluation object to improve the quality and service of airlines.

E. METHOD OF RESEARCH

The research methods used in this research are:

1. Data Collection Method
   a. Interview
      At this stage an interview process is conducted on the airlines about the information required for this research.
   b. Study of Literature
      At this stage, searching information that support the research e.g Data Mining, Naive Bayes Classification, Prediction System from books, journals, e-books, and websites.
2. Software Development Method
   Software development method used in this research is Waterfall Model. Software development starts from system analysis process, system design, system encoding, system testing, and system implementation and maintenance.

II. RESEARCH METHOD

A. DEFINITION OF PREDICTION

Prediction is a process of estimating something that is most likely to happen in the future based on past and present information, so that the error (the difference between something that happens and the expected result) can be minimized.

There are 2 types of prediction techniques:

1. Qualitative Predictions
   Qualitative predictions are based on qualitative data in the past. Qualitative methods are used if past data of predicted variables are not present, not sufficient.
2. Quantitative Prediction
   Quantitative predictions are based on quantitative data in the past. The predicted results depend on the method used in the prediction.

(Herdianto, 2013)

B. CLASSIFICATION

Classification is a work of assessing data objects to include them in a class of available classes. In the classification there are two main step: first, the development of the model as a prototype to be stored as memory and second, the use of the model to do the introduction / classification / prediction on another object data to be known in the class where the data object is stored.
(Prasetyo, 2012)

C. NAÏVE BAYES CLASSIFICATION

Naive Bayes is a probabilistic classifier based on Bayes Rule of conditional probability. Naive Bayes uses the probability of classifying new instances. The workings of Naive Bayes itself is to look for the greatest opportunity number of possible classification, by looking at the frequency of each classification in the training data. The Bayesian classification is based on the Bayes theorem, which is derived from the name of the British mathematician and prebysterian minister Thomas Bayes (1702-1761).

The equations of the bayes theorem are:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

Explanation :

X	= appearance of overall characteristics
H	= appearance of characteristics in the class
P(H)	= probability hypothesis H (prior probability)
P(X)	= probability of X
P(H|X)	= probability ofX based H condition (posterior probability)
P(X|H)	= probability of H based X condition

Naive bayes is a simplification of the Bayes theorem. The equation of Naive Bayes are:

$$P(H|X) = P(X|H) \times P(X)$$

(Septari, 2014)

D. WATERFALL MODEL

The software development model first introduced by Royce in 1970 comes from the adaptation of hardware development, because at that time there was no other software development methodology. The existence of the flow from one stage to another, this model is referred to as the waterfall model. The waterfall model is a plan-based development where all activities must be planned and scheduled before starting a job.

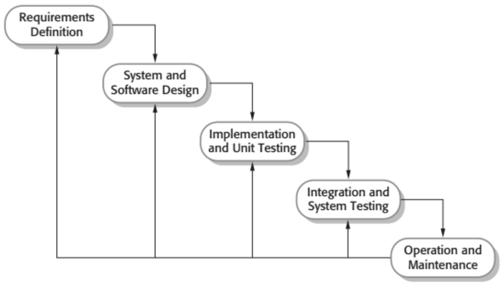
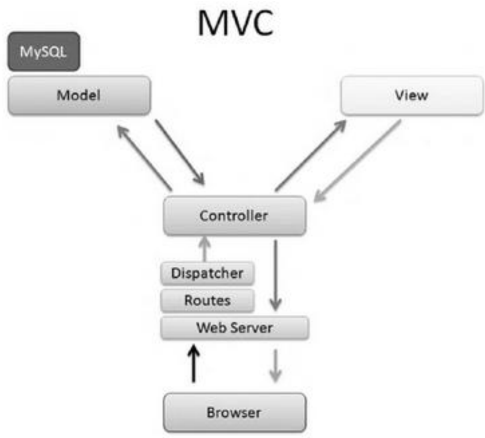


Figure 2.1 Diagram Model Waterfall (Sommerville, 2011)

E. PHP CODEIGNITER

Codeigniter is an open source application of framework with MVC model (Model, View, Controller) to build dynamic website using PHP. Codeigniter makes it easy for web developers or developers to create web apps quickly and easily than making from scratch.



Gambar 2.2 MVC Model (Supono & Putratama, 2016)

F. CONFUSION MATRIX

Confusion matrix is a table to measure the performance of classification algorithms or classifier. In confusion matrix there are some terms commonly used:

1. *True Positive* (TP) :prediction data is true and the fact is true.
2. *True Negative* (TN) :prediction data is true and the fact is false.
3. *False Positive* (FP) :prediction data is false and the fact is true.

4. *False Negative* (FN) :prediction data is false and the fact is false.

Table 2.1CalculationConfusion *Matrix*(Markham, 2014)

Nama	Rumus
Accuracy	(TP + TN) / N
Error Rate	(FP + FN) / N
TP Rate	TP / (TP + FN)
FP Rate	FP / (TN + FP)
Specificity	TN / (TN + FP)
Precision	TP / (FP + TP)
Prevalence	*Actual Positive* / N

III. RESULTS AND ANALYSIS

A. ANALYSIS

1. Input requirement
   a. Data Flight History
   Data flight historycontains departure date, departure time, flight number, aircraft registration, aircraft type, origin, destination, ontime status.
   b. Data Flight
   Data flight is the data that will be predicted for the punctuality of departure. This data contains departure date, departure time, flight number, aircraft registration, aircraft type, origin, destination.
2. Process requirement
   a. Process data flight
   Processing flight data contains about data processing both flight history data and flight data to be predicted.
   b. Process flight prediction
   The process of predicting flights contains about the calculation of the possibility of a flight on time or delay.
3. Output requirement
   a. OTP prediction information
   The information contains percentageof on time performance based prediction data.
   b. Prediction information
   The information contains detail of prediction such as departure date, departure time, flight number, aircraft registration, aircraft type, origin, destination, on-time status.

B. PREDICTION SYSTEM ANALYSIS

The application to be developed is an application to predict the punctuality of flight departure using naive bayes algorithm. Input for app is test data (flight data 2-8 January 2017) and train data (flight data January 2016). Test data will be calculated the probability with reference history data. Flowchart described :

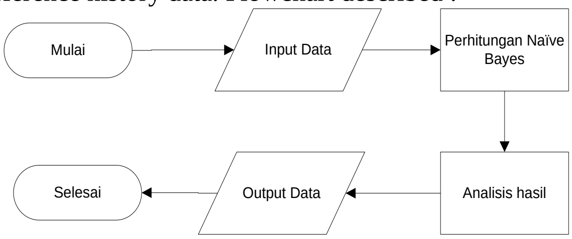


Figure 3.1Flow Algorithm Naive Bayes

C. BUSINESS ANALYSIS OF PREDICTION APPLICATION

Here is the business process analysis:

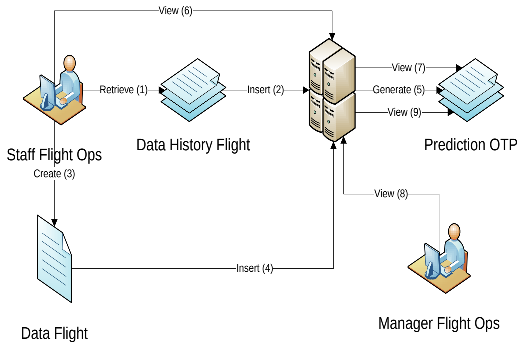


Figure 3.2Business Process Analysis

D. ALGORITHM DESIGN

Here is flow of calculating punctuality flight departure:

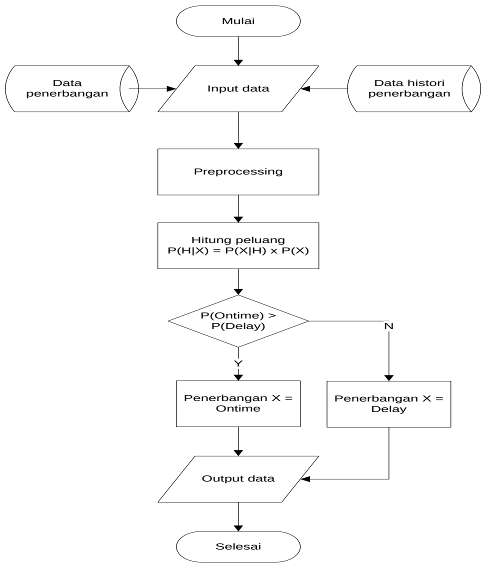


Figure3.3Flow Prediction Calculation

E. ALGORITHM SIMULATION

To do the classification it takes the training data and test data as input in this algorithm. Training data is flight history data and test data is data to be searched for flight departure time.

Table 3.1 Training data Set

DepDay	FlightNum	Origin	Destination	AcType	AcReg	Ontime
Senin	894	CGK	UPG	A330	PKGPT	YES
Senin	648	JOG	DPS	B738	PKGFQ	NO
Selasa	654	CGK	PLM	B738	PKGMU	NO
Selasa	400	SUB	JOG	B738	PKGMY	YES
Rabu	500	CGK	PNK	A330	PKGMA	YES
Rabu	604	KNO	BTH	B738	PKGFY	YES
Kamis	100	DPS	JOG	A330	PKGEH	YES

Table 3.2 Test Data Set

DepDay	FlightNum	Origin	Destination	AcType	AcReg	Ontime
Senin	202	CGK	JOG	A330	PKGPT	???

In the calculation process, a unique class is not included in the calculation. The deleted class is FlightNum. Here are the calculation steps:

1. Counts class ontime
   a. P(Y=Yes) = 5/7
   b. P(Y=No) = 2/7
2. Counts same case based class ontime
   a. P(X | Y=Yes)
      1. P(Depday = Senin | Y = Yes) = 1/5
      2. P(Origin = CGK | Y = Yes) = 2/5
      3. P(Destination = JOG | Y = Yes) = 2/5
      4. P(AcType = A330 | Y = Yes) = 3/5
   b. P(X | Y=No)
      1. P(Depday = Senin | Y = No) = 0/2
      2. P(Origin = CGK | Y = No) = 1/2
      3. P(Destination = JOG | Y = No) = 0/2
      4. P(AcType = A330 | Y = No) = 0/2
3. Multiply all variable:
   a. P(X | Y = Yes) = 5/7 x 1/5 x 2/5 x 2/5 x 3/5 = 0,01371
   b. P(X | Y = No) = 2/7 x 0/2 x 1/2 x 0/2 x 0/2 = 0
4. Compare result multiply:
   The calculation of the Ontime "Yes" class with the Ontime "No" class indicates that the Ontime "Yes" class has a larger value than the Ontime "No" class. Then it can be deduced that:
   Class Ontime = Yes

F. USE CASE APPLICATION

Here is use case of application:

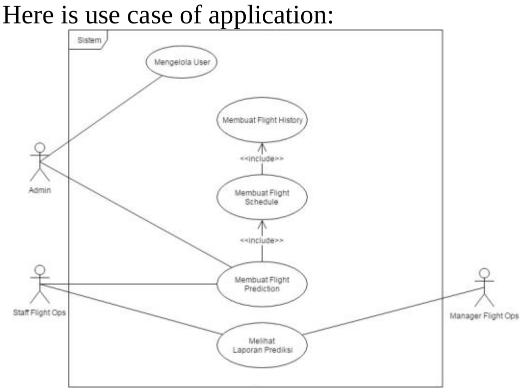


Figure3.4 Use Case Application

The actor's definition of the above use case are :

Table 3.3 Use Case Actor Definition

Actor	Description
Admin	The person assigned as administrator of the app and has full access rights to the app.
Staff Flight Ops	An application user whose permissions are limited only to modules related to the flight prediction function.
Manager Flight Ops	It is an application user who only has permissions to view the flight prediction report.

Here is description of use case:

Table3.4 Use Case Description

Use Case	Description
Manage User	Is a user management process that can enter in the application. There are functions to view the user, add users, edit users, and delete the user.
Create Flight History	It is an flight history management process where there is a function to view history, add history, edit history, and delete history.
Create Flight Schedule	It is an upcoming flight schedule management process. There is a function to view schedules, add schedules, edit schedules, and delete schedules.
Create Flight Prediction	Is a process to predict flights from flight schedules.
ViewPrediction Report	Is a process to see the results of the predicted flight schedule that has been done.

G. DATA MODEL

Here is information about data model application
:

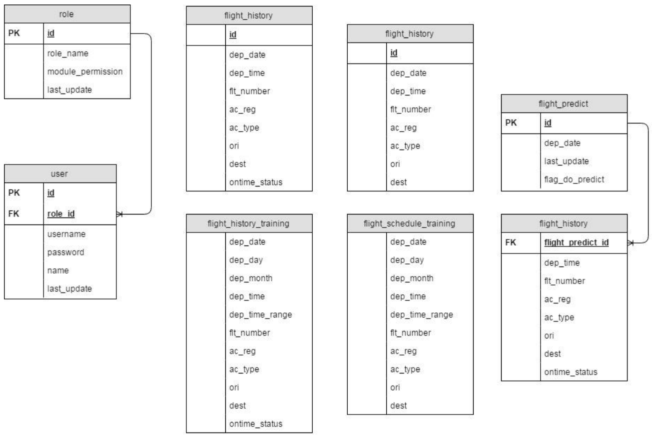


Figure 3.1 Data Model Application

H. IMPLEMENTATION RESULT

Here is implementation of flight historypage :

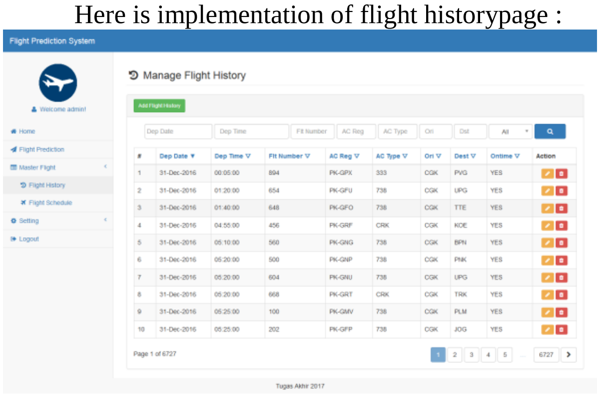


Figure 3.6Implementation Manage Flight History Page

Here is implementation of flight schedulepage :

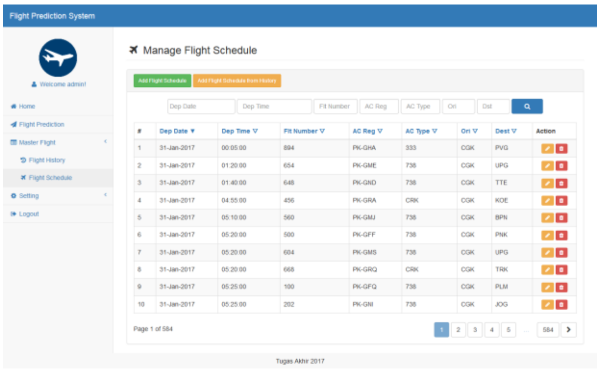


Figure 3.7Implementation Flight Schedule Page

Implementationof result prediction page :

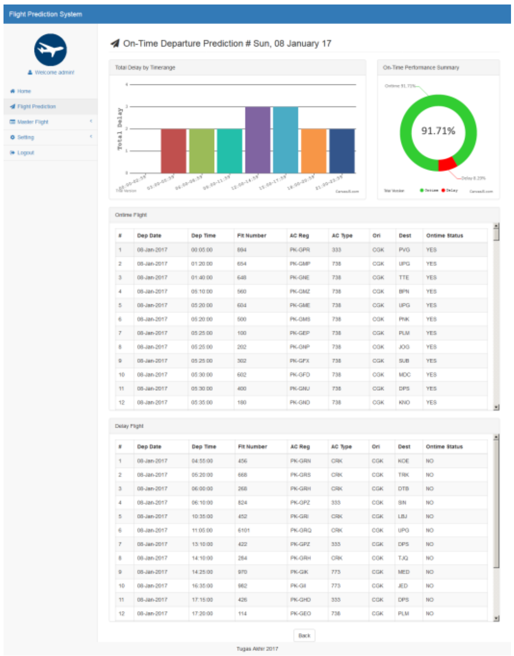


Figure 3.8 Implementation Flight Prediction Result Page

I. WHITE-BOX TESTING RESULT

White-box testing is done by checking the logic in the program code. Step to do white-box testing is create a flowchart of program code then mapped to flowgraph. From the flowgraph will be analyzed cyclomatic complexity and connection matrix.

Flowchart algorithm :

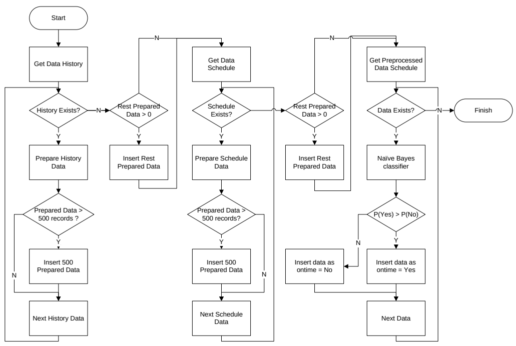


Figure3.9 Flowchart Algorithm

Flow graph based flow chart :

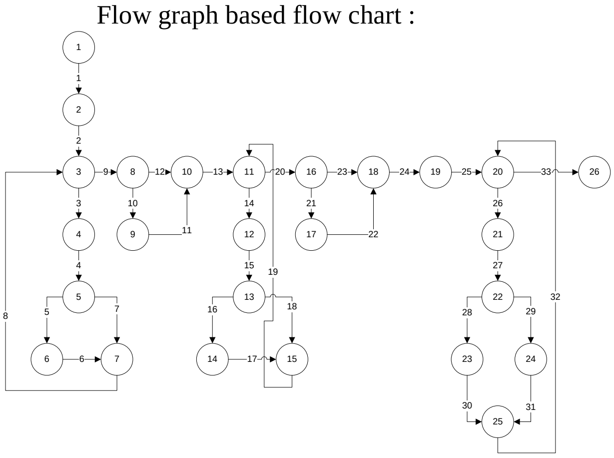


Figure3.10Flowgraph

Cyclomatic complexity formula :

V(G) = E – N + 2

Based flow graph above then it can be conducted :

E (Edge) = 33

N (Node) = 26

V(G) = 33 – 26 + 2 = 9

Here is matrix connection :

N\CN	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	Koneksi
1		1																									0
2			1																								0
3				1				1																			1
4					1																						0
5						1	1																				1
6							1																				0
7			1																								0
8									1	1																	1
9										1																	0
10											1																0
11												1				1											1
12													1														0
13														1	1												1
14															1												0
15											1																0
16																	1	1									1
17																		1									0
18																			1								0
19																				1							0
20																					1					1	1
21																						1					0
22																							1	1			1
23																									1		0
24																									1		0
25																				1							0
																										Total	8

Figure3.11Matrix Connection

Cyclomatic complexity formula :

V(G) = P + 1

Based matrix conection above then it can be conducted :

P (Node Connection) = 8

V(G) = 8 + 1 = 9

J. CONFUSION MATRIX RESULT

The data tested is the flight on 2-8 January 2017. Information obtained from the prediction results compared with the actual data is as follows:

```
description    total
-----------  --------
TP               1019
TN                 25
FP                239
FN                 69
```

Figure3.12ComparationPrediction Data and Actual Data

Based on the total data above it can be calculated as follows:

Table3.5Calculation of Confusion Matrix

Accuracy	(TP+TN)/N	77,22%
Misclassification Rate	(FP+FN)/N	22,78%
TP Rate	TP/(TP+FN)	93,66%
FP Rate	FP/(FP+TN)	90,53%
Specificity	TN/(FP+TN)	9,47%
Precision	TP/(TP+FP)	81,00%
Prevalence	(TP+FN)/N	80,47%

Based on the results of calculations that have been done, obtained correctly classified or accuracy of 77.22%. Correctly classified is the percentage of the number of classes predicted according to the actual class. With true positive rate (sensitivity) accuracy of 93.66%, true negative rate (specificity) of 9.47%, positive predictive value (precision) of 81.00%, accuracy of 77.22%. Sensitivity is used to compare the number of true positive to the number of positive records whereas specifity, precision is the ratio of true negative numbers to the number of negative records. Accuracy that produces values in the range 70% - 80% indicates that the naive bayes algorithm classified to the fair classification.

IV. CONCLUSION

Based on a list of theory, analysis, design, implementation and testing software that has been done, it can be concluded that:

1. Classification of naive bayes has been successful in predicting the punctuality of flight departure and naive bayes algorithm can be implemented with an accuracy of 77.22% and classified as fair classification or sufficiently categorized.
2. Based on the calculation of complexity and the connection matrix cocluded that the value of V(G) for both is equal, ie 9. This explains that there is no logical error in the program code.

ACKNOWLEDGEMENTS

This research was supported by head of research center Dr Devi Fitrianah,.S.Kom,.MTI who have given us foundation to finish our research in their place.

BIBLIOGRAPHY

1) Direktorat Jenderal Perhubungan Udara. (2016, Feb 3). *On Time Performance 15 Maskapai Berjadwal Periode Juli-Desember 2015 Sebesar 77,16%*. Dipetik Maret 27, 2017, dari Kementerian Perhubungan Republik Indonesia: http://dephub.go.id/post/read/on-time-performance-15-maskapai-berjadwal-periode-juli-desember-2015-sebesar-77,16

2) Bahaweres, R. B., Alaydrus, M., & Wahab, A. (2012). ANALISIS KINERJA VOIP CLIENT SIPDROID DENGAN MODUL ENKRIPSI, *2012*(Snati), 15–16.

3) Dwi, B. (2015, Oktober 16). *Fungsional dari Notepad++*. Dipetik Mei 17, 2017, dari Bayu Dwi Arta Pratama: https://bayudwiarta.wordpress.com/2014/10/16/fungsional-dari-notepad/

4) Han, J., & Kamber, M. (2012). *Data Mining : Concepts and Techniques.* San Fransisco: Morgan Kauffman Publishers.

5) Herdianto. (2013). *Prediksi Kerusakan Motor Induksi Menggunakan Metode Jaringan Saraf Tiruan Backpropagation.* Medan: Universitas Sumatera Utara.

6) Markham, K. (2014, Maret 25). *Simple guide to confusion matrix terminology*. Dipetik Mei 13, 2017, dari Dataschool: http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

7) Permana, I. (2012, Juni 15). *Contoh UML Parkir*. Dipetik Mei 27, 2017, dari http://ikper.blogspot.co.id/2012/06/contoh-uml-parkir.html

8) Prasetyo, E. (2012). *Data Mining Konsep dan Aplikasi Menggunakan MATLAB.* Yogyakarta: Andi.

9) Pressman, R. S. (2012). *Rekayasa Perangkat Lunak.* Jakarta: Andi.

10) Ridwan, M., Suryono, H., & Sarosa, D. (2012). Penerapan Data Mining Untuk evaluasi Kinerja Akademik Mahasiswa menggunakan Algoritma Naive Bayes. 59.

11) Santosh, K. (2014). Computational Intelligence in Data Mining - Volume 2. Dalam *Classification of Heart Disease Using Naive Bayes and Genetic* (hal. 269-282).

12) Saputra R, A. (2014). Komparasi Algoritma Klasifikasi Data Mining Untuk Memprediksi Penyakit Tuberculosis (TB): Studi Kasus Puskesmas Karawang Sukabumi. *Seminar Nasional Inovasi dan Tren*, (hal. 1-8).

13) Septari, A. (2014). *APLIKASI PREDIKSI PEMINATAN SISWA SMAN 8 BANDUNG DENGAN MENGGUNAKAN KLASIFIKASI DATA MINING ALGORITMA NAIVE BAYES* . Bandung.

14) Sommerville, I. (2011). *Sommerville - Software Engineering 9th Edition.* Massachusetts: Pearson Education.

15) Supono, & Putratama, V. (2016). *Pemrograman Web dengan Menggunakan PHP dan Framework Codeigniter*. Bandung: Deepublish.

16) Syarif, A., Abouaissa, A., Idoumghar, L., Sari, R. F., & Pascal Lorenz. (2014). Performance analysis of evolutionary multi-objective based approach for deployment of wireless sensor network with the presence of fixed obstacles. *IEEE*. https://doi.org/10.1109/GLOCOM.2014.7036775

17) Ventura, B. (2017, Februari 1). *BPS: Jumlah Penumpang Pesawat Tahun 2016 Capai 95,2 Juta*. Dipetik Mei 13, 2017, dari SINDONEWS.com: https://ekbis.sindonews.com/read/1176254/33/bps-jumlah-penumpang-pesawat-tahun-2016-capai-952-juta-1485965749

18) Wikipedia. (t.thn.). *Bandar Udara Soekarno Hatta*. Dipetik Mei 13, 2017, dari Wikipedia Indonesia: https://id.wikipedia.org/wiki/Bandar_Udara_Internasional_Soekarno-Hatta

19) Wikipedia. (2016, Agustus 3). *On-time performance*. Dipetik Mei 10, 2017, dari Wikipedia: https://en.wikipedia.org/wiki/On-time_performance