Available online at: http://proceeding.rsfpress.com/index.php/ ess/index
LPPM UPN "Veteran" Yogyakarta Conference Series
Proceeding on Engineering and Science Series (ESS)
Volume 1 Number 1 (2020): 328-339

Modeling of Crude Oil Types Classification Using the Naive Bayes Classifier Method

Harry Budiharjo Sulistyarso, Dyah Ayu Irawati, Joko Pamungkas, Indah Widiyaningsih

Universitas Pembangunan Nasional Veteran Yogyakarta
E-mail address harry_hb@upnyk.ac.id, E-mail address:dyah.ayu.irawati.@upnyk.ac.id, E-mail address: joko.pamungkas@upnyk.ac.id, E-mail address : indahwidiyaningsih@upnyk.ac.id

Abstract

This research is part of previous research regarding the prediction of petroleum's physical properties to help the user get the prediction value of crude oil's physical properties from field test data, which was carried out from Enhanced Oil Recovery Research Laboratory, Petroleum Engineering UPN "Veteran" Yogyakarta. The field data that is measured in the laboratory that has been done is by adding biosurfactants and increasing the temperature. Various steps have been taken to reduce crude oil's viscosity value so that it can be diluted could flow. It is necessary to calculate the viscosity of crude oil in this process to determine the extent of the viscosity level as expected by adding biosurfactants or increasing the temperature that has been carried out in the EOR process. Naïve Bayes Classifier is used to classify oil data into three categories: light oil, medium oil, and heavy oil, based on the prediction of the viscosity value. The Naïve Bayes Classifier is a robust algorithm for performing machine learning-based predictive modeling that applies the Bayes theorem. This predictive modeling for the physical properties of crude oil was built using the Python programming language and the PyQt5 library to build desktop-based applications. The classification of oil has arrived at labeling the prediction results of crude oil's viscosity into three categories, namely light oil, medium oil, and heavy oil. The test results with testing data produce accurate data; the predicted value is within the specified range.

Keywords: Crude oil, EOR, Naïve Bayes, Prediction, Viscosity

I. INTRODUCTION

In Indonesia, the consumption of fuel oil is still very high, while the source of oil is decreasing. One way to meet the demand for fuel oil is by increasing oil production through the Enhanced Oil Recovery process, which is able to reactivate oil wells that are no longer used, and the remaining oil is categorized as heavy crude oil. In the research that has been done, the EOR process is carried out by adding biosurfactant gradually and increasing the temperature from low to a maximum of 80C in order to reduce the viscosity level so that crude oil can be streamed, then the generated data from the laboratory trials are used to perform predictive modeling of petroleum's physical

properties so that the expected values of physical properties can be known quickly, accurately and can be carried out in all conditions. The values of the physical properties of processed crude oil are viscosity, Interfacial Tension (IFT), and density. (Sulistyarso, HB et al., 2020).

The determination of the viscosity and pour point is carried out to ensure the flow characteristics of crude oil at low temperatures. However, there are some general relationships for crude oil composition that can be derived from the pour point and viscosity data. Generally, the lower the pour point of crude oil, the higher the aromatics, and the higher the pour point, the more paraffinic. Viscosity is usually determined at different temperatures (for example, between 25 ° C and 100 ° C) by measuring the time for the volume of liquid to flow under gravity through a calibrated glass capillary viscometer. (ASTM D445-11).

In this study, from experiments conducted in the laboratory, crude oil is categorized into 3 types based on its viscosity value:

a. Light crude oil, containing low levels of metals and sulfur, is light in color and is dilute (low viscosity 1-10 cp).
b. Medium crude oil, containing moderate levels of metals and sulfur, is a bit dark in color and is a bit thin (medium viscosity 10-50 cp)
c. Heavy crude oil (heavy crude oil), containing high levels of metals and sulfur, has a high viscosity, so it must be heated to melt (high viscosity> 50 cp)

Classification of crude oil and labeling of the final results of the forecast needs to be done to make it easier for users to know the final yield of crude oil obtained from predictive modeling of the physical properties of petroleum. Naïve Bayes Classifier is used to predict the classification of oil data into three categories, namely light oil, medium oil, and heavy oil. The Naïve Bayes Classifier is a powerful algorithm for performing machine learning-based predictive modeling that applies the Bayes theorem. Predictive modeling for the physical properties of petroleum is built using the Python programming language and the PyQt5 library to build desktop-based applications.

II. LITERATURE REVIEW

This part discusses the previous related study and the Naive Bayes method.

II.1. Previous Research

Research related to the Naïve Bayes method referred to in this study is research conducted by Vural and Gök (2017). In this study, it was stated that the perpetrators' most likely criminal incidents could be predicted through the criminal records that are owned by the police. The dataset used in this study was developed from an artificial dataset due to the difficulty in obtaining crime rate data from the police due to confidentiality considerations. The variables contained in the dataset include the date and location of the incident, type of crime, criminal ID, and acquaintance. An acquaintance is a suspect whose name is directly involved in the incident or is an indirect acquaintance of the criminal. The proposed system was tested for criminal prediction problems using cross-validation, and the experimental results showed that the proposed system gave a high score of 78.05% in finding criminal suspects.

The next research referred to in this study is the research conducted by Sarkar S and Sriram Ram S (2001). In this study, the Bayes theorem is compared with the Composite Attributes and Induced Decision Tree models in predicting a bank's health. The comparison of the three prediction techniques is that all three have an adequate level of accuracy, namely an average of 91.5%.

Further research on the implementation of Naïve Bayes is a study conducted by Ayu Irawati et al. (2018). In this study, an expert system mobile application using the Naïve Bayes method has been built, which can help rabbit breeders detect disease in rabbits through visible symptoms.

II.2. Classification with Naïve Bayes

Classification is the process of finding functions and models that can distinguish or explain concepts or data classes to estimate the unknown class of an object.

In the usual classification process, two processes must be carried out, namely (Nugroho & Subanar, 2013):

II.2.1. Training Process.

In this process, training data sets or sample data with known labels or attributes will build the model.

II.2.2. Testing process

The testing process is done to determine the accuracy of the model made in the training process, so data is built called testing data to classify the labels.

Naïve Bayes is a simple probability classification that calculates a set of probabilities by summing the frequencies and value combinations from a given dataset. The algorithm uses the Bayes theorem and assumes that all the attributes are independent or interdependent, given the class variable's value. Naïve Bayes is also defined as a classification using the probability and statistical method proposed by the British scientist Thomas Bayes, which predicts future opportunities based on previous experiences (Saleh, 2015).

The equation of the Bayes theorem can be seen below:

$$P(H|X) = \frac{P(X|H).P(H)}{P(H)} \qquad (1)$$

Where:

X: data with an unknown class

H: the data hypothesis uses a specific class

P (H | X): the probability of hypothesis H under condition X (partial probability)

P (H): probability hypothesis H (prior probability)

P (X | H): probability X based on the conditions of hypothesis H

P (X): probability H

II.3. System Overview

The application built in the previous research is the application of the prediction of petroleum's properties, which are viscosity, IFT, and density. The applications discussed in this article are part of these applications, namely a feature to classify types of petroleum from the viscosity prediction results, which are carried out into three categories, namely light oil, medium oil, and heavy oil.

III. RESEARCH METHODOLOGY

The steps taken to solve the problem are described in Figure 1. The waterfall model suggests a sequential approach to software development that begins with customer specification of requirements and progresses through planning, modeling, construction, and deployment, culminating in the ongoing support of the completed software. The explanation is as follows:

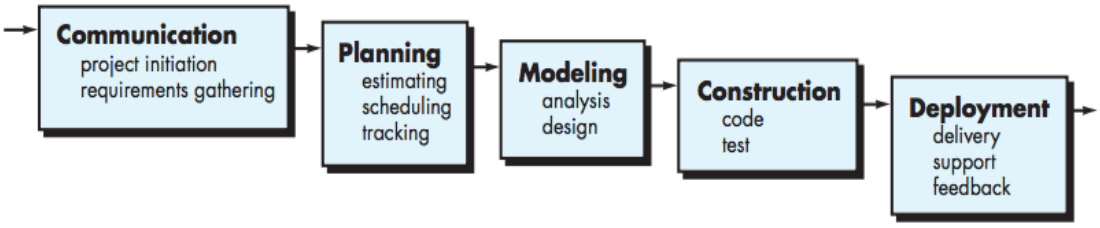

Figure 1. Research Methodology Flow with the Waterfall Model (Pressman, S. Roger.)

III.1. COMMUNICATION

This phase is the initial phase of the research. In this phase, the required data is collected. The data needed in this study are as follows:

a. The need for petroleum sample data from the field
b. Data from laboratory test results regarding the petroleum's physical properties from the samples obtained are then used as a dataset in the prediction application for the type of petroleum, based on the viscosity to be built.

There is a need for predictive modeling of the type of petroleum that can be automated from the dataset of laboratory trials that are owned. The prediction application model that will be built will later be made using the Naïve Bayes method. The application will have the ability to calculate the probability of each variable, then compare the probability between variables and provide prediction results in the form of labels for types of petroleum, namely light oil, medium oil, heavy oil.

Table 1 shows the data obtained from laboratory trials, and then the data is used as a raw data dataset in this study.

Tabel 1. Initiation Dataset

Biosurfactant	Temp °C	Viscosity (CP)
0%	30	5,48
0%	40	3,36
0%	50	2,88
0%	60	2,08
0%	70	1,1
0%	80	1
5%	30	3,29
5%	40	1,16

5%	50	0,79
5%	60	0,91
5%	70	0,58
5%	80	0,41
10%	30	1,55
10%	40	1,21
10%	50	1,03
10%	60	0,94
10%	70	0,87
10%	80	0,7

III.2. Planning

Before designing the system, analysis is carried out first of the existing data to determine all the necessary support so that the research is completed correctly. This stage will produce a user requirements document or data related to user needs in making software, including plans to be carried out, estimated work times, tools needed, division of work.

III.3. Modelling

The design process will translate the requirements into a software design that can be estimated before the program is created. This process focuses on: data structures, software architecture, interface representations, and procedural details (algorithms).

As the initial description of the modeling application that will be built, the following is the design:

III.3.1. System Flowchart Design

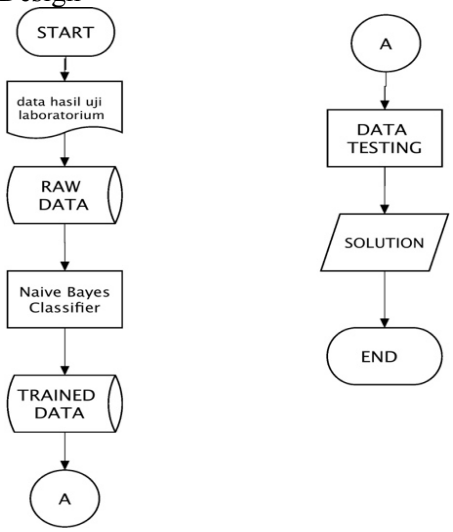

Figure 2. Flowchart of How Applications Work

At the beginning of making the application, laboratory test result data must first be digitized into a database. The dataset will later become raw data, which will be classified with the Naïve Bayes Classifier and trained so that the results will later become trained data.
Because the data available are numerical, the steps for Naïve Bayes are depicted in Figure 3.

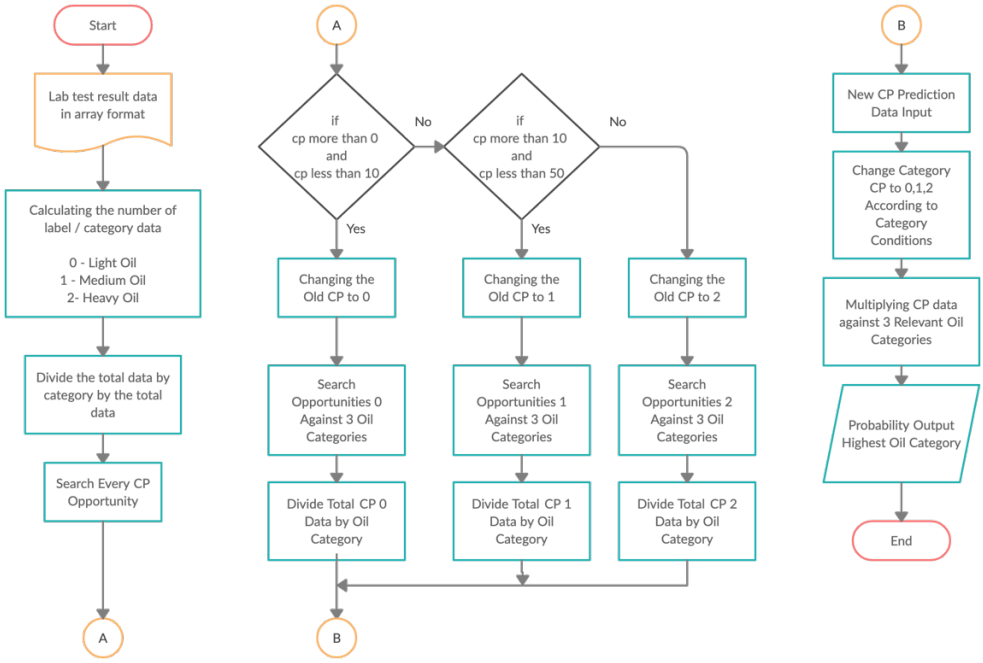

Figure 3. Step by Step of Naïve Bayes Classification

III.3.2. Mock Up Design
Figure 4 illustrates the features of a petroleum classification based on the viscosity resulting from the application's predicted calculations.

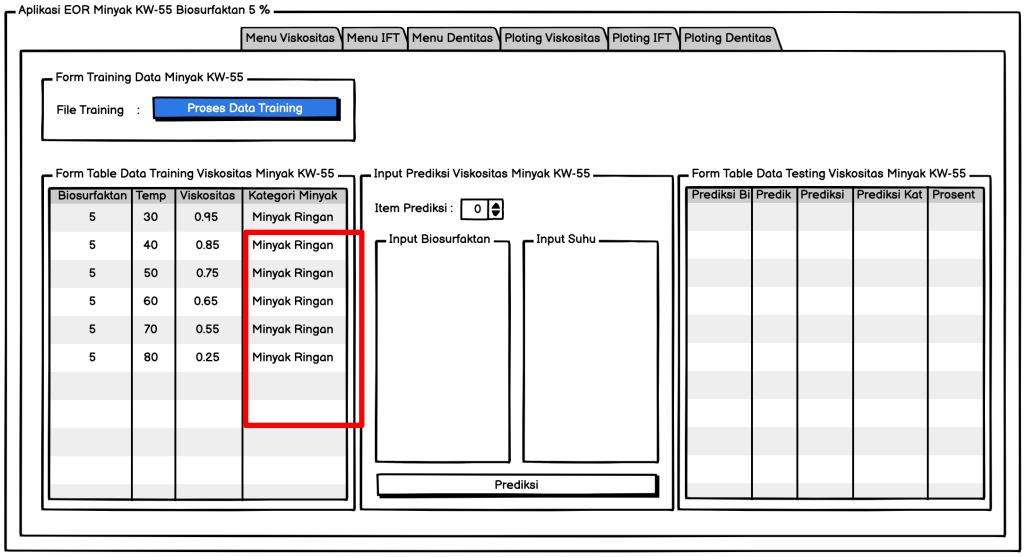

Figure 4. Application Mock Up

III.4. Construction

This section describes the steps for the classification of petroleum-based on Naïve Bayes. Table 2 contains the distribution of viscosity ranges and labeling. Table 3 shows data from the field mapped against three oil categories based on viscosity. The viscosity data in table 2 results from predictive modeling applications for the physical properties of petroleum.

Table 2. Division of viscosity range and labeling.

Viscosity Range	Classification	Label
0 – 10	Light Oil	0
11 – 50	Medium Oil	1
> 50	Heavy il	2

Table 3. Data from viscosity mapping results for three categories of petroleum with Naïve Bayes

Viscosity	Category Label
5,48	Light Oil
3,36	Light Oil
2,88	Light Oil
2,08	Light Oil
1,1	Light Oil
1	Light Oil
3,29	Light Oil
1,16	Light Oil
0,79	Light Oil
0,91	Light Oil
0,58	Light Oil
0,41	Light Oil
1,55	Light Oil
1,21	Light Oil
1,03	Light Oil
0,94	Light Oil
0,87	Light Oil
0,7	Light Oil
12,3	Medium Oil
11	Medium Oil
55	Heavy Oil
56	Heavy oil
52,2	Heavy Oil

Search for probability labels on the lab test data table

$$\mathrm{P}(Light\ Oil) = \frac{Amount\ of\ light\ oil\ data}{Total\ data}$$
$$\mathrm{P}(Light\ Oil) = \frac{18}{23}$$
$$\mathrm{P}(Light\ Oil) = 0{,}782608696$$

$$(medium\ Oil) = \frac{Amount\ of\ medium\ oil\ data}{Total\ data}$$
$$\mathrm{P}(Medium\ Oil) = \frac{2}{23}$$
$$\mathrm{P}(Medium\ Oil) = 0{,}086956522$$

$$(heavy\ Oil) = \frac{Amount\ of\ heavy\ oil\ data}{Total\ data}$$
$$(heavy\ Oil) = \frac{3}{23}$$
$$(heavy\ Oil) = 0{,}130434783$$

Simplified search for probabilities against viscosity data

$$(category\ Label\ 0|light\ oil) = \frac{amount\ of\ Category\ 0\ \text{data}}{amount\ of\ light\ oil\ data}$$
$$\mathrm{P}(Category\ Label\ 0|light\ oil) = \frac{18}{18}$$
$$\mathrm{P}(Category\ Label\ 0|light\ oil) = 1$$

$$(category\ Label\ 0|medium\ oil) = \frac{amount\ of\ Category\ 0\ \text{data}}{amount\ of\ medium\ oil\ data}$$
$$\mathrm{P}(Category\ Label\ 0|medium\ oil) = \frac{0}{2}$$
$$\mathrm{P}(Category\ Label\ 0|medium\ oil) = 0$$

$$(category\ Label\ 0|heavy\ oil) = \frac{amount\ of\ Category\ 0\ \text{data}}{amount\ of\ heavy\ oil\ data}$$
$$(category\ Label\ 0|heavy\ oil) = \frac{0}{3}$$
$$(category\ Label\ 0|heavy\ oil) = 0$$

$$(category\ Label\ 1|light\ oil) = \frac{amount\ of\ Category\ 1\ \text{data}}{amount\ of\ light\ oil\ data}$$
$$\mathrm{P}(Category\ Label\ 1|light\ oil) = \frac{0}{18}$$
$$(category\ Label\ 1|light\ oil) = 0$$

$$\mathrm{P}(Category\ Label\ 1|medium\ oil) = \frac{amount\ of\ Category\ 1\ \mathrm{data}}{amount\ of\ medium\ oil\ data}$$
$$\mathrm{P}(Category\ Label\ 1|medium\ oil) = \frac{2}{2}$$
$$\mathrm{P}(Category\ Label\ 1|medium\ oil) = 1$$

$$\mathrm{P}(Category\ Label\ 1|heavy\ oil) = \frac{amount\ of\ Category\ 1\ \mathrm{data}}{amount\ of\ heavy\ oil\ data}$$
$$\mathrm{P}(Category\ Label\ 1|heavy\ oil) = \frac{0}{3}$$
$$\mathrm{P}(Category\ Label\ 1|heavy\ oil) = 0$$

$$\mathrm{P}(Category\ Label\ 2|light\ oil) = \frac{amount\ of\ Category\ 2\ \mathrm{data}}{amount\ of\ light\ oil\ data}$$
$$\mathrm{P}(Category\ Label\ 2|light\ oil) = \frac{0}{18}$$
$$\mathrm{P}(Category\ Label\ 2|light\ oil) = 0$$

$$\mathrm{P}(Category\ Label\ 2|medium\ oil) = \frac{amount\ of\ Category\ 2\ \mathrm{data}}{amount\ of\ medium\ oil\ data}$$
$$\mathrm{P}(Category\ Label\ 2|medium\ oil) = \frac{0}{2}$$
$$\mathrm{P}(Category\ Label\ 2|medium\ oil) = 0$$

$$\mathrm{P}(Category\ Label\ 2|heavy\ oil) = \frac{amount\ of\ Category\ 2\ \mathrm{data}}{amount\ of\ heavy\ oil\ data}$$
$$\mathrm{P}(Category\ Label\ 2|heavy\ oil) = \frac{3}{3}$$
$$\mathrm{P}(Category\ Label\ 2|heavy\ oil) = 1$$

Table 4 The results of the calculation of the label/category of oil properties on the presented data

Category	Data amount	Total Data	Result
Light Oil	18	23	0,782608696
Medium oil	2	23	0,086956522
Heavy Oil	3	23	0,130434783

Table 5. Testing data for experiments to see the properties of oil-based on viscosity

Category	CP: 5.48 – Label (0)	result	Output
Light Oil	1	1	Light Oil
Medium Oil	0	0	
Heavy Oil	0	0	

$$P(0 \mid 5.48) = \frac{amount\ of\ Category\ 0\ \text{data}}{amount\ of\ light\ oil\ data}$$
$$P(0 \mid 5.48) = \frac{18}{18}$$
$$P(0 \mid 5.48) = 1$$

$$P(0 \mid 5.48) = \frac{amount\ of\ Category\ 0\ \text{data}}{amount\ of\ medium\ oil\ data}$$
$$P(0 \mid 5.48) = \frac{0}{2}$$
$$P(0 \mid 5.48) = 0$$

$$P(0 \mid 5.48) = \frac{amount\ of\ Category\ 0\ \text{data}}{amount\ of\ heavy\ oil\ data}$$
$$P(0 \mid 5.48) = \frac{0}{3}$$
$$P(0 \mid 5.48) = 0$$

III.5 Deployment

This section presents the results of applications that have been successfully built and used. Figures 5 and 6 show that the application has been able to categorize the oil according to the viscosity values that appear.

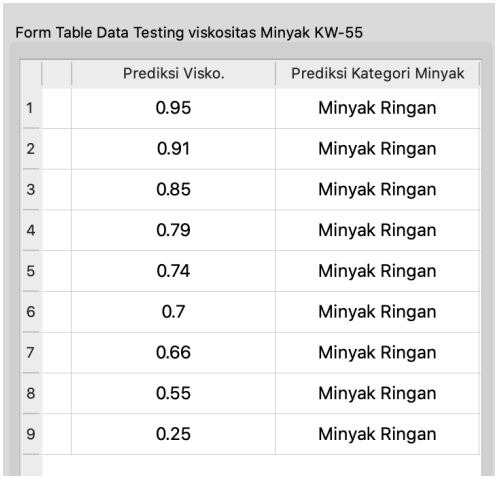

Figure 5. Categorization of Petroleum, based on the predicted viscosity value

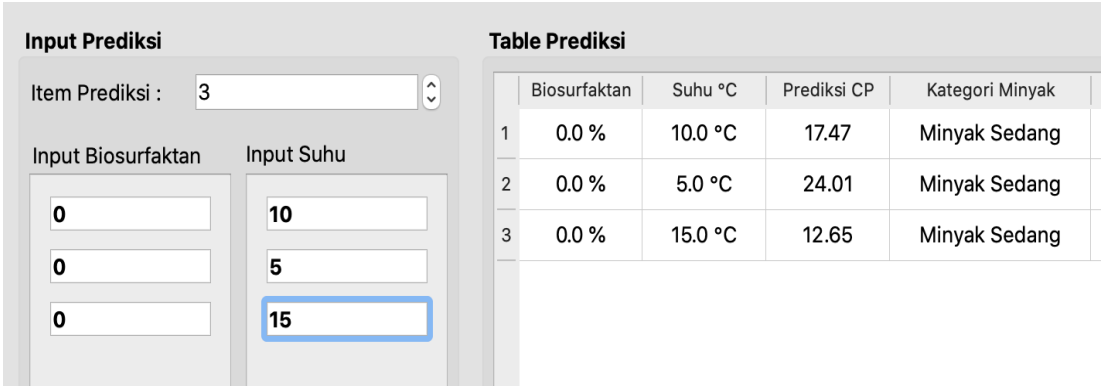

Figure 6. Categorization of Petroleum, based on the predicted viscosity value

IV. FINDING AND DISCUSSION

In this study, the validity test of application calculations was carried out using the same value as the initial dataset.

To test application validity, use the range of conditions on the vitality label, which is stated in the table below.

Table 6. Viscosity range and labeling.

Viscosity Range	Classification	Label
0 – 10	Light Oil	0
11 – 50	Medium Oil	1
> 50	Heavy il	2

The viscosity prediction results can be seen by the label and viscosity of the oil; when the oil's viscosity properties do not match the prediction range of the validity of truth, the label is considered false/invalid.

V. CONCLUSION AND FURTHER RESEARCH

The classification of crude oil and labeling the final results of the prediction can be done. The results obtained to follow the specified range make it easier for users to know the final yield of crude oil obtained from the prediction modeling of petroleum's physical properties.

VI. REFERENCES

ASTM D445-11, Standard Test Method for Kinematic Viscosity of Transparent and Opaque Liquids (and Calculation of Dynamic Viscosity). Designation 71/1/97. *ASTM International, 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959, United States*. pp.2-3.

Ayu Irawati D. et al., 2018. "Development of Android-based Rabbit Disease Expert System." International Journal of Engineering and Technology. 7(4.44). DOI: 10.14419/ijet.v7i4.44.26868. pp. 82-87

Nugroho A., Subanar. 2013. "Klasifikasi Naïve Bayes untuk Prediksi Kelahiran pada Data Ibu Hamil". Jurnal Berkala MIPA Vol 23, No.3(2013) ISSN:0215-9309. Fakultas MIPA UGM. Pp

297-308.

Pressman, S. Roger. Software Engineering: A Practitioner's Approach, Eight Edition. 2015. McGraw Hill. The USA. p. 42

Saleh, A. 2015. Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. Citec Journal, Vol. 2, No. 3, ISSN: 2354-5771. pp 207-217.

Sarkar S dan Sriram Ram S. 2001. Bayesian Models for Early Warning Bank Failures. Management Science Journal. 47(11). pp 1457–1475.

Sulistyarso HB., Pamungkas, Joko, Gusmawarni, Sri Rahayu, Widyaningsih, Indah, Kurnia, Rafli Arie. 2020. "*Application of Bio-surfactants as an Effort to Enhanced Oil Recovery (EOR) in Kawengan Oil Field*," AIP Conference Proceedings 2245, 090018. pp. 1-5.

Vural MS dan Gök M. 2017. Criminal prediction using Naïve Bayes Theory. ACM Journal Neural Computing and Applications. 28(9) pp 2581–2592. DOI 10.1007/s00521-016-2205-z.