CHAPTER 4
ANALYSIS AND DESIGN

4.1 Analysis

This analysis section discusses how to implement the Naive Bayes Algorithm. Here's how to implement it :

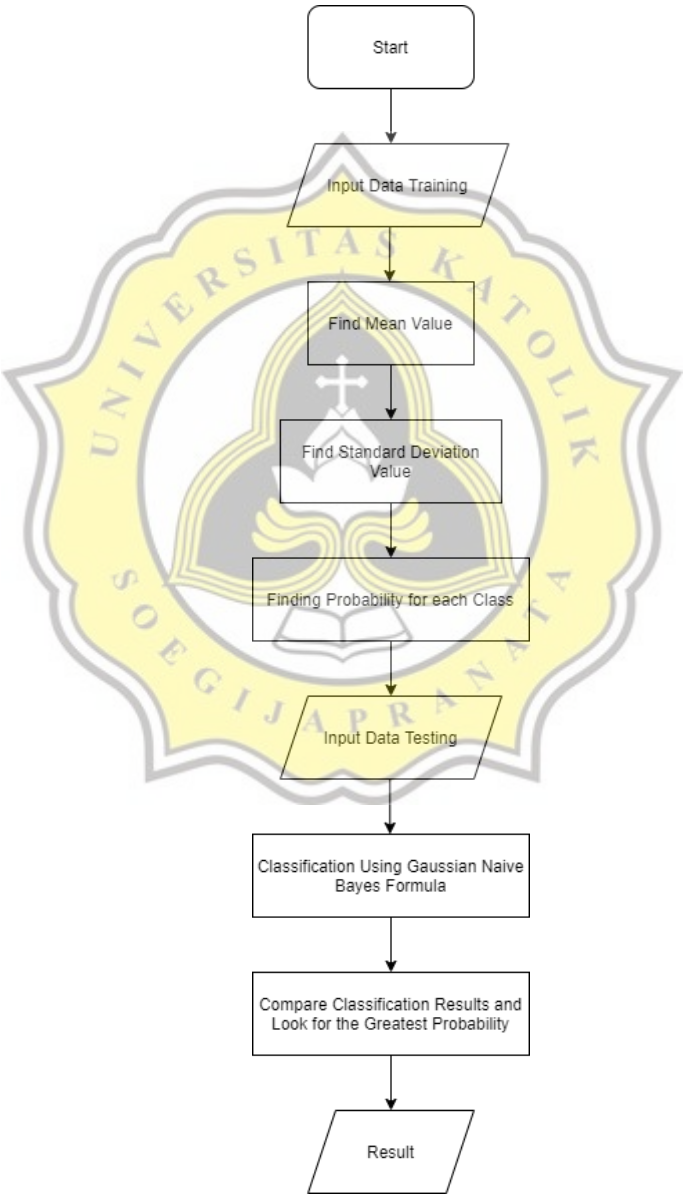

Illustration 4.1: Implementation of the Naive Bayes Algorithm

The first step in the Naive Bayes implementation process is to enter and read the data. The training data has been processed so that it can be used in program calculations with a change in data type, namely for the parameter Gender, the male data is changed to number 1 and for the female data it is changed to number 2. For the guarantee parameter, K data is changed to number 3 and for NK data changed to number 4. The training data used are as follows:

Table 4.1: Training Data Example Table

No	Gender	Submission	Salary	Period	Interest	Guarantee	Class
1	2	5000000	2117500	10	2.5	3	Approved
2	1	7000000	4298432	6	2.5	3	Approved
3	2	7500000	3637500	6	2.5	3	Approved
4	2	5000000	1685000	10	2.5	3	Approved
5	1	5000000	1685000	12	2.5	3	Approved
6	2	5000000	2167500	10	2.5	3	Approved
7	1	7000000	3082500	10	2.5	3	Approved
8	2	1000000	2200000	1	4	3	Approved
9	2	3000000	5000000	12	2.5	4	Approved
10	1	3500000	3100000	12	2.5	4	Approved
11	2	5000000	3400000	12	2.5	3	Disapproved
12	2	5000000	2600000	12	2.5	3	Disappoved
13	2	50000000	8000000	24	2	4	Disapproved
14	1	50000000	4000000	24	2	4	Disapproved
15	2	60000000	3250000	48	2	4	Disapproved

The next step is to find the Mean value of the training data using the formula:

$$X = \frac{\sum Xi}{N}$$

Illustration 4.2: Mean Formula

Information : X = Mean

Xi = data value to i

N = total data

This stage is carried out to find the mean value of each class. This value is obtained by adding up each data in the class then dividing by the number of data in that class. The calculation method is as follows:

Mean calculation of the Gender parameter :

Approved Class = (2+1+2+2+1+2+1+2+2+1) / 10
= 16 / 10
= 1.6

Disapproved Class = (2+2+2+1+2) / 5
= 9 / 5
= 1.8

Mean calculation of the Submission parameter :

Approved Class = (5000000+7000000+7500000+5000000+
5000000+5000000+7000000+1000000+
3000000+3500000) / 10
= 49000000 / 10
= 4900000

Disapproved Class = (5000000+5000000+50000000+
50000000+60000000) / 5
= 170000000 / 5
= 34000000

Mean calculation of the Salary parameter :

Approved class = (2117500+4298432+3637500+1685000+
1685000+2167500+3082500+2200000+
5000000+3100000) / 10
= 28973432 / 10
= 2897343.2

Disapproved Class = (3400000+2600000+8000000+4000000+
3250000) / 5
= 21250000 / 5
= 4250000

Mean calculation of the Period parameter :

Approved Class = (10+6+6+10+12+10+10+1+12+12) / 10
= 89 / 10
= 8.9

Disapproved Class = (12+12+24+24+48) / 5
= 120 / 5
= 24

Mean calculation of the Interest parameter :

Approved Class = (2.5+2.5+2.5+2.5+2.5+2.5+2.5+4+2.5+
2.5) / 10
= 26.5 / 10
= 2.65

Disapproved Class = (2.5+2.5+2+2+2) / 5

= 11 / 5

= 2.2

Mean calculation of the Guarantee parameter :

Approved Class = (3+3+3+3+3+3+3+3+4+4) / 10

= 32 / 10

= 3.2

Disapproved Class = (3+3+4+4+4) / 5

= 18 / 5

= 3.6

The following is a summary table for calculating the mean of each class:

Table 4.2: Mean Calculation Summary Table

Class	Gender	Submission	Salary	Period	Interest	Guarantee
Approved	1.6	4900000	2897343.2	8.9	2.65	3.2
Disapproved	1.8	34000000	4250000	24	2.2	3.6

After getting the Mean value, the next step is to find the standard deviation value for each class with the formula :

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

Illustration 4.3: Standard Deviation Formula

Information : σ = Standard Deviation

xi = data value to i

μ = Mean value

N = total data minus 1 (n-1)

At this stage, the standard deviation value is sought by subtracting the average value of the data class by subtracting each data value and then adding up all the results. Here's how to calculate it :

Standard Deviation calculation of the Gender parameter :

Approved Class

$\sum(x_i - \mu)^2$ = ((2-1.6)^2)+((1-1.6)^2)+((2-1.6)^2)+((2-1.6)^2)+ ((1-1.6)^2)+((2-1.6)^2)+((1-1.6)^2)+((2-1.6)^2)+ ((2-1.6)^2)+((1-1.6)^2) = 2.4

$\frac{\sum(x_i - \mu)^2}{N}$ = 2.4 / 9 = 0.267

$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ = $\sqrt{0.267}$ = 0.5163

Disapproved Class

$\sum(x_i - \mu)^2$ = ((2-1.8)^2)+((2-1.8)^2)+((2-1.8)^2)+((1-1.8)^2)+ ((2-1.8)^2) = 0.8

$\frac{\sum(x_i - \mu)^2}{N}$ = 0.8 / 4 = 0.2

$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ $= \sqrt{0.2} = 0.4472$

Standard Deviation calculation of the Submission parameter :

Approved Class

$\sum(x_i - \mu)^2$ = ((5000000-4900000)^2)+((7000000-4900000)^2)+
((7500000-4900000)^2)+((5000000-4900000)^2)+
((5000000-4900000)^2)+((5000000-4900000)^2)+
((7000000-4900000)^2)+((1000000-4900000)^2)+
((3000000-4900000)^2)+((3500000-4900000)^2)
= 36400000000000

$\frac{\sum(x_i - \mu)^2}{N}$ = 36400000000000 / 9 = 4044444444444.44

$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ $= \sqrt{4044444444444.44} = 2011080.4172$

Disapproved Class

$\sum(x_i - \mu)^2$ = ((5000000-34000000)^2)+((5000000-34000000)^2)+
((50000000-34000000)^2)+((50000000-34000000)^2)+
((60000000-34000000)^2)
= 2870000000000000

$\frac{\sum(x_i - \mu)^2}{N}$ = 2870000000000000 / 4 = 717500000000000

$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ $= \sqrt{717500000000000} = 26786190.4719$

Standard Deviation calculation of the Salary parameter :

Approved Class

$\sum(x_i - \mu)^2$ = ((2117500-2897343.2)^2)+((4298432-2897343.2)^2)+
((3637500-2897343.2)^2)+((1685000-2897343.2)^2)+
((1685000-2897343.2)^2)+((2167500-2897343.2)^2)+
((3082500-2897343.2)^2)+((2200000-2897343.2)^2)+
((5000000-2897343.2)^2)+((3100000-2897343.2)^2)
= 11574066500000

$\frac{\sum(x_i - \mu)^2}{N}$ = 11574066500000 / 9 = 1286007390000

$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ = $\sqrt{1286007390000}$ = 1134022.6566

Disapproved Class

$\sum(x_i - \mu)^2$ = ((3400000-4250000)^2)+((2600000-4250000)^2)+
((8000000-4250000)^2)+((4000000-4250000)^2)+
((3250000-4250000)^2)
= 18570000000000

$\frac{\sum(x_i - \mu)^2}{N}$ = 18570000000000 / 4 = 4642500000000

$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ = $\sqrt{4642500000000}$ = 2154646.1426

Standard Deviation calculation of the Period parameter :

Approved Class

$\sum(x_i - \mu)^2$ = ((10-8.9)^2)+((6-8.9)^2)+((6-8.9)^2)+((10-8.9)^2)+ ((12-8.9)^2)+((10-8.9)^2)+((10-8.9)^2)+((1-8.9)^2)+ ((12-8.9)^2)+((12-8.9)^2) = 112.9

$\frac{\sum(x_i - \mu)^2}{N}$ = 112.9 / 9 = 12.54444444

$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ = $\sqrt{12.54444444}$ = 3.5418

Disapproved Class

$\sum(x_i - \mu)^2$ = ((12-24)^2)+((12-24)^2)+((24-24)^2)+((24-24)^2)+ ((48-24)^2) = 864

$\frac{\sum(x_i - \mu)^2}{N}$ = 864 / 4 = 216

$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ = $\sqrt{216}$ = 14.6969

Standard Deviation calculation of the Interest parameter :

Approved Class

$\sum(x_i - \mu)^2$ = ((2.5-2.65)^2)+((2.5-2.65)^2)+((2.5-2.65)^2)+ ((2.5-2.65)^2)+((2.5-2.65)^2)+((2.5-2.65)^2)+ ((2.5-2.65)^2)+((4-2.65)^2)+((2.5-2.65)^2)+ ((2.5-2.65)^2) = 2025

$\frac{\sum(x_i - \mu)^2}{N}$ = 2025 / 9 = 0.225

$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ = $\sqrt{0.225}$ = 0.4743

Disapproved Class

$\sum(x_i - \mu)^2$ = ((2.5-2.2)^2)+((2.5-2.2)^2)+((2-2.2)^2)+((2-2.2)^2)+ ((2-2.2)^2) = 0.3

$\frac{\sum(x_i - \mu)^2}{N}$ = 0.3 / 4 = 0.075

$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ = $\sqrt{0.075}$ = 0.2738

Standard Deviation calculation of the Guarantee parameter :

Approved Class

$\sum(x_i - \mu)^2$ = ((3-3.2)^2)+((3-3.2)^2)+((3-3.2)^2)+((3-3.2)^2)+ ((3-3.2)^2)+((3-3.2)^2)+((3-3.2)^2)+((3-3.2)^2)+ ((4-3.2)^2)+((4-3.2)^2) = 1.6

$\frac{\sum(x_i - \mu)^2}{N}$ = 1.6 / 9 = 0.1777777778

$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$ = $\sqrt{0.1777777778}$ = 0.4216

Disapproved Class

$\sum(x_i - \mu)^2$ = ((3-3.6)^2)+((3-3.6)^2)+((4-3.6)^2)+((4-3.6)^2)+ ((4-3.6)^2) = 1.2

$$\frac{\sum(x_i - \mu)^2}{N} \quad = 1.2 / 4 = 0.3$$

$$= \sqrt{\frac{\sum(x_i - \mu)^2}{N}} \quad = \sqrt{0.3} = 0.5477$$

The following table summarizes the calculation of Standard Deviation for each class :

Table 4.3: Standard Deviation Summary Calculation Table

Class	Gender	Submission	Salary	Period	Interest	Guarantee
Approved	0.51	2011080.41	1134022.65	3.54	0.47	0.42
Disapproved	0.44	26786190.47	2154646.14	14.69	0.27	0.54

The next step is to find the probability of each class by dividing the total number of data for a class by the total number of data. Here's an example of the calculation :

Approved class probability

= total data of the Approved class / total amount of data

= 10 / 15 = 0.6667

Disapproved class probability

= total data of the Disapproved class / total amount of data

= 5 / 15 = 0.3333

Table 4.4: Summary Table of Probability Calculation for each Class

Class	Probability
Approved	0.6667
Disapproved	0.3333

After getting the probability value for each class, the next step is to determine the testing data. The following is an example of testing data :

Table 4.5: Sample Data Testing Table

No	Gender	Submission	Salary	Period	Interest	Guarantee	Class
1	2	7000000	6500000	10	2.5	3	?

The next step is to find the final probability value by applying the Gaussian Naive Bayes algorithm with the following formula :

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Illustration 4.4: Gaussian Naive Bayes Formula

Information :

P : Probability

X : data to x

σ = Standard Deviation

μ = Mean value

The following is an example of calculating the final probability using the Gaussian Naive Bayes formula :

Final Probability Calculation for Gender parameter :

Approved Class

$$f(x) = \frac{1}{0.5163\sqrt{2*3.14}} e^{-\frac{(2-1.6)^2}{2*0.5163^2}}$$

$$f(x) = \frac{1}{1.2938} e^{-\frac{0.16}{0.5331}} = 0.5725$$

Disapproved Class

$$f(x)=\frac{1}{0.4472\sqrt{2*3.14}}e^{-\frac{(2-1.8)^2}{2*0.4472^2}}$$

$$f(x)=\frac{1}{1.1206}e^{-\frac{0.04}{0.3999}} = 0.8074$$

Final Probability Calculation for Submission parameter :

Approved Class

$$f(x)=\frac{1}{2011080.4171\sqrt{2*3.14}}e^{-\frac{(7000000-4900000)^2}{2*2011080.4171^2}}$$

$$f(x)=\frac{1}{5039753.08}e^{-\frac{4410000000000}{8088888890000}} = 1.1503$$

Disapproved Class

$$f(x)=\frac{1}{26786190.4719\sqrt{2*3.14}}e^{-\frac{(7000000-34000000)^2}{2*26786190.4719^2}}$$

$$f(x)=\frac{1}{67126000.92}e^{-\frac{729000000000000}{1435000000000000}} = 8.9635$$

Final Probability Calculation for Salary parameter :

Approved Class

$$f(x)=\frac{1}{1134022.6566\sqrt{2*3.14}}e^{-\frac{(6500000-2897343.2)^2}{2*1134022.6566^2}}$$

$$f(x)=\frac{1}{2841852.632}e^{-\frac{12979136000000}{2572014770000}} = 2.2637$$

Disapproved Class

$$f(x)=\frac{1}{2154646.1426\sqrt{2*3.14}}e^{-\frac{(6500000-4250000)^2}{2*2154646.1426^2}}$$

$$f(x)=\frac{1}{5399527.757}e^{-\frac{5062500000000}{9285000000000}}=1.0736$$

Final Probability Calculation for Period parameter :

Approved Class

$$f(x)=\frac{1}{3.5418\sqrt{2*3.14}}e^{-\frac{(10-8.9)^2}{2*3.5418^2}}$$

$$f(x)=\frac{1}{8.8757}e^{-\frac{1.21}{25.0886}}=0.1073$$

Disapproved Class

$$f(x)=\frac{1}{14.6969\sqrt{2*3.14}}e^{-\frac{(10-24)^2}{2*14.6969^2}}$$

$$f(x)=\frac{1}{36.8303}e^{-\frac{196}{431.9977}}=0.0172$$

Final Probability Calculation for Interest parameter :

Approved Class

$$f(x)=\frac{1}{0.4743\sqrt{2*3.14}}e^{-\frac{(2.5-2.65)^2}{2*0.4743^2}}$$

$$f(x)=\frac{1}{1.1885}e^{-\frac{0.0225}{0.4499}}=0.8003$$

Disapproved Class

$$f(x)=\frac{1}{0.2738\sqrt{2*3.14}}e^{-\frac{(2.5-2.2)^2}{2*0.2738^2}}$$

$$f(x)=\frac{1}{0.6861}e^{-\frac{0.09}{0.1499}}=0.7995$$

Final Probability Calculation for Guarantee parameter :

Approved Class

$$f(x)=\frac{1}{0.4216\sqrt{2*3.14}}e^{-\frac{(3-3.2)^2}{2*0.4216^2}}$$

$$f(x)=\frac{1}{0.4216\sqrt{2*3.14}}e^{-\frac{(3-3.2)^2}{2*0.4216^2}}=0.8457$$

Disapproved Class

$$f(x)=\frac{1}{0.5477\sqrt{2*3.14}}e^{-\frac{(3-3.6)^2}{2*0.5477^2}}$$

$$f(x)=\frac{1}{1.3725}e^{-\frac{0.36}{0.5999}}=0.3998$$

The following is a summary table of the Final Probability values :

Table 4.6: Final Probability value table

Class	Gender	Submission	Salary	Period	Interest	Guarantee
Approved	0.5725	1.1503	2.2637	0.1073	0.8003	0.8457
Disapproved	0.8074	8.9635	1.0736	0.0172	0.7995	0.3998

After knowing the final probability calculation results, the next step is to calculate the classification value for each class by multiplying all the final probabilities by the probability value for each class.

Approved Class Classification Value

= (Probability of Gender x Probability of Submission x Probability of Salary x Probability of Period x Probability of Interest x Probability of Guarantee x Class Probability) Approved

= 0.5725 x 1.1503 x 2.2637 x 0.1073 x 0.8003 x 0.8457 x 0.6667 = 0.0721

Disapproved Class Classification Value

= (Probability of Gender x Probability of Submission x Probability of Salary x Probability of Period x Probability of Interest x Probability of Guarantee x Class Probability) Disapproved

= 0.8074 x 8.9635 x 1.0736 x 0.0172 x 0.7995 x 0.3998 x 0.3333 = 0.0142

The next step is to compare the classification results of each class. The highest classification value will be chosen to be the final result of the classification process. From the results of the above calculations it can be concluded that the training data is included in the Approved classification.

Table 4.7: Final Data Testing Result Table

No	Gender	Submission	Salary	Period	Interest	Guarantee	Class
1	2	7000000	6500000	10	2.5	3	Approved

4.2 Design

Flowchart

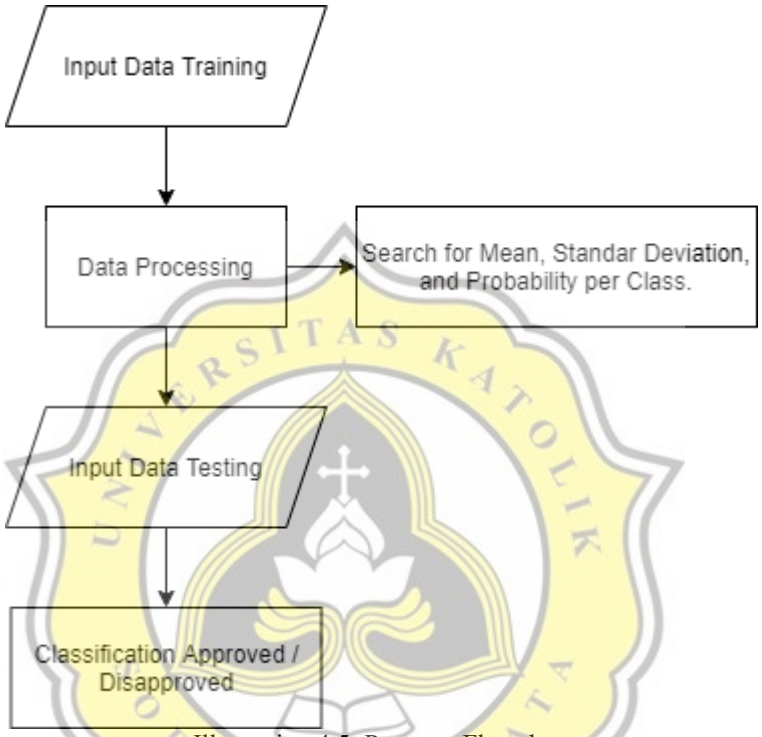

Illustration 4.5: Program Flowchart

In accordance with the flowchart above, the first step that must be taken is to input the training data into the program. The data used is data in csv format. Furthermore, the training data is processed by looking for the mean value, standard deviation, and also the probability of each class in the training data. After getting the results, the next step is to input Data Testing which will be calculated with the results of the Data Training calculations using the Naive Bayes algorithm. After the calculation process using the Naive Bayes Classifier method is complete, the classification results will be obtained whether the prospective customer is in the Approved / Disapproved category.