CHAPTER 4

ANALYSIS AND DESIGN

4.1. Collecting Data

Sample data taken is data from Kaggle, with share link: https://www.kaggle.com/sjleshrac/airlines-customer-satisfaction/. This data is provided by the airline with the airline name invistico, the existing data set consists of the details of the customers who have flown with them, then with the feedback it is recorded as source data. With the following column information, the first column explains about age, then the second column explains the flight distance traveled, then the next column is the location of the plane's entry gate, then the next column explains about wifi service, then the next column describes entertainment in airplane flights, then the next column explains then explained about online support during flights, then the next column explained about the ease of booking airline tickets online, then the next column was in-flight service, then column nine was for in-flight cleanliness, then for the tenth column explained about the delay in departure in how many minutes, then for the eleventh column describes the delay in arrival in how many minutes, then the last column satisfaction column with values 1 and 0 whose purpose is to see what is the difference between satisfied and dissatisfied.

Table 1: Data Sample Table

No	Age	Flight Distance	Gate Location	Inflight wifi service	Inflight entertainment	Online support	Ease of Online booking	On-board service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	Satisfaction
1	7	1241	4	1	1	1	1	3	4	0	0	0
2	47	2464	3	0	2	2	3	4	3	310	305	1
3	46	1863	4	1	1	1	1	4	4	20	22	0
4	47	2464	3	0	2	2	3	4	3	310	305	1
5	46	1863	4	1	1	1	1	4	4	20	22	0
6	15	2138	3	2	0	2	2	3	4	0	0	1
7	37	1701	4	5	1	5	5	3	3	1	0	0
8	60	623	3	3	4	3	1	1	1	0	0	1
9	70	354	3	4	3	4	2	2	2	0	0	1
10	21	1255	4	1	1	1	1	4	4	86	75	0

4.2.1 SVM Algorithm Steps

1. Split Data

Split Dataset To separate training and testing data The training data used is starting from column number one to number eight, while numbers nine to ten are testing data. The purpose of distinguishing data separation is to facilitate calculations in the algorithm process.

A. Training Data:

No	Age	Flight Distance	Gate Location	Inflight wifi service	Inflight entertainment	Online support	Ease of Online booking	On-board service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	Satisfaction
1	7	1241	4	1	1	1	1	3	4	0	0	0
2	47	2464	3	0	2	2	3	4	3	310	305	1
3	46	1863	4	1	1	1	1	4	4	20	22	0
4	47	2464	3	0	2	2	3	4	3	310	305	1
5	46	1863	4	1	1	1	1	4	4	20	22	0
6	15	2138	3	2	0	2	2	3	4	0	0	1

7	37	1701	4	5	1	5	5	3	3	1	0	0
8	60	623	3	3	4	3	1	1	1	0	0	1

B. Testing Data:

No	Age	Flight Distance	Gate Location	Inflight WIFI service	Inflight entertainment	Online support	Ease of Online booking	On-board service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	Satisfaction
9	70	354	3	4	3	4	2	2	2	0	0	1
10	21	1255	4	1	1	1	1	4	4	86	75	0

2. Implementation to Formula

Next is the calculation of the SVM algorithm with the data set entered by the following formula below, then the data is processed with the aim of getting the W value, the W value is the weight/attribute value of each variable, there are eleven attribute values.

Table 2: Formula

Formula Used		
1.	$1/2\|\| W\|\|^2 = 1/2\ (W_1^2 + W_2^2 + w_3^2 + W_3^2 + W_4^2 \ldots. W_{11}^2)$	With $Y_1\ (w.xi + b) > 1,\ i = 1$
2.	$Y_1\ (w_1.x_1 + w_2.x_2 + w_3.x_3 + w_4.x_4 \ldots\ldots. + W_{11}^2) \geq +1$	To Satisfy
3.	$Y2\ (w_1.x_1 + w_2.x_2 + w_3.x_3 + w_4.x_4 \ldots\ldots. + W_{11}^2) \geq -1$	For Dissatisfied (Or 0)

A. From the training data, the following equation can be obtained:

1.	$(W_1.7)\ (W_2.1241)\ (W_3.4)\ (W_4.1)$ (W5.1) (W6.1) (W7.1) (W8.3) (W9.4) (W10.0) (W11.0) - b ≥ - 1
2.	$(W_1.47)\ (W_2.2464)\ (W_3.3)\ (W_4.0)$ (W5.2) (W6.2) (W7.3) (W8.4) (W9.3) (W10.310) (W11.305) + b≥ + 1
3.	$(W_1.46)\ (W_2.1863)\ (W_3.4)\ (W_4.1)$ (W5.1) (W6.1) (W7.1) (W8.4) (W9.4) (W10.20) (W11.22) - b ≥ - 1
4.	$(W_1.47)\ (W_2.2464)\ (W_3.4)\ (W_4.1)$ (W5.1) (W6.1) (W7.1) (W8.4) (W9.4) (W10.20) (W11.0) + b ≥ + 1
5.	$(W_1.46)\ (W_2.1863)\ (W_3.4)\ (W_4.1)$ (W5.1) (W6.1) (W7.1) (W8.4) (W9.4) (W10.20) (W11.0) - b ≥ - 1
6.	$(W_1.15)\ (W_2.2138)\ (W_3.3)\ (W_4.2)$ (W5.0) (W6.2) (W7.2) (W8.3) (W9.4) (W10.0) (W11.0) + b ≥ + 1
7.	$(W_1.37)\ (W_2.1701)\ (W_3.4)\ (W_4.5)$ (W5.1) (W6.5) (W7.5) (W8.3) (W9.3) (W10.0) (W11.0) - b ≥ - 1
8.	$(W_1.15)\ (W_2.623)\ (W_3.3)\ (W_4.3)$ (W5.4) (W6.3) (W7.1) (W8.1) (W9.1) (W10.0) (W11.0) + b ≥ + 1

B. From the Testing data, the following equation can be obtained:

1.	$(W_1.70)\ (W_2.354)\ (W_3.3)\ (W_4.4)$ (W5.3) (W6.4) (W7.2) (W8.2) (W9.2) (W10.0) (W11.0) + b ≥ + 1
2.	$(W_1.21)\ (W_2.1255)\ (W_3.4)\ (W_4.1)$ (W5.1) (W6.1) (W7.1) (W8.4) (W9.86) (W10.75) (W11.0) - b ≥ - 1

3. Finding the Value of Each Attribute

Then after successfully getting the equations in each data branch, the next step is to get the value of each attribute value, perform linear calculations on each equation so that the value of each attribute value can be found by eliminating the following eliminations.

A. Elimination Calculation:

1.	$-7w_1 - 1241w_2 - 4w_3 - w_4 - w_5 - w_6 - w_7 - 3w_8 - 4w_9 - w_{10} - w_{11} - b = -1$ (1) $47w_1 + 2464\ w_2 + 3w_3 + w_4 + 2w_5 + 2w_6 + 3w_7 + 4w_8 + 3w_9 + 310w_{10} + 305w_{11} + b = 1$ (2)
	$-54w_1 - 2588w_2 - 7w_3 - 2w_4 - 3w_5 - 3w_6 - 4w_7 - 7w_8 - 7w_0 - 311w_{10} - 306w_{11} - 2b = -2$ (9)
2.	$47w_1 + 2464\ w_2 + 3w_3 + w_4 + 2w_5 + 2w_6 + 3w_7 + 4w_8 + 3w_9 + 310w_{10} + 305w_{11} + b = 1$ (2) $-46w_1 - 1863w_2 - 4w_3 - w_4 - w_5 - w_6 - w_7 - 4w_8 - 4w_9 9 - 20w_{10} - 22w_{11} - b = -1$ (3)
	$93w_1 + 4327w_2 + 7w_3 + 2w_4 + 3w_5 + 3w_6 + 4w_7 + 8w_8 + 7w_9 + 330_{w10} + 327w_{11} + 2b = 2$ (10)
3.	$-46w_1 - 1863w_2 - 4w_3 - w_4 - w_5 - w_6 - w_7 - 4w_8 - 4w_9 9 - 20w_{10} - 22w_{11} - b = -1$ (3) $47w_1 + 2464\ w_2 + 3w_3 + w_4 + 2w_5 + 2w_6 + 3w_7 + 4w_8 + 3w_9 + 310w_{10} + 305w_{11} + b = 1$ (4)
	$-93w_1 - 4327w_2 - 7w_3 - 2w_4 - 3w_5 - 3w_6 - 4_{w7} - 8w_8 - 7w_9 - 330w_{10} - 327w11 - 2b = -2$ (11)

Then do these steps until the last equation in the same way, namely subtracting between equations

B. Elimination of W Values Using Equation

1.	$78w_1 + 3478w_2 + 2w_8 + 38w_{10} + 42w_{11} + b = 1$ (16) $-37w_1 - 3478w_2 - 4w_8 - 32w_{10} - 12w_{11} + 8b = 8$(17)
	$450w_1 - 6w_8 + 70w_{10} - 30w_{11} - 7b = -7$ (18)
2.	$-37w_1 - 3478w_2 - 4w_8 - 32w_{10} - 12w_{11} + 8b = 8$(17) $-450w_1 - 6w_8 + 70w_{10} - 30w_{11} - 7b = -7$ (18)
	$2232w_1 + 24w_8 + 192w_{10} + 72w_{11} + 48b = 48$ $1800w_1 + 24w_8 + 280w_{10} + 120\ w_{11} + 28b = 28$
	$432\ w_1 - 88w_{10} - 48w_{11} - 20b = -20$ (19)

Then after succeeding to get the last equation that is used to find the value of each attribute value with the elimination technique, then the next step is to find the value with the substitution technique

C. Finding the Value of W with the equation

1.	$-432w_1 - 48w_{11} - 20b = -20$ (19) $-69840w_1 - 456w_2 - 784b = -784b$ (20) $992w_1 + 9120b = 9120$ $3.352.320w_1 + 37632b = 37632$	
		$-3155321\ w_1 = -28512$ $W_1 = -0{,}00903$
2.	$-372w_1 - 3478w_2 - 8b = 8$ $-372(-0{,}00903) - 3472w_2 = 8$	
		$W_2 = -1{,}335$
3.	$-37w_1 - 170w_2 - 4w_3 - b = -1$ $37\ (-0{,}00903) - 170(-1{,}335) - 4w_3 - b = 1$	
		$b = 1{,}57$

Then it is done repeatedly to find the value of each attribute until you find the value of each attribute as below:

W1	-0,00903	W7	-560,47
W2	-1,335	W8	- 63,74
W3	56,571	W9	-4,406
W4	49.498,44	W10	-146,541
W5	-373,04	W11	-3,56
W6	782,382	B	1,57

4. Conclusion Using Data Testing

If you have found the value of each attribute, what you do is do a trial test by entering it into the testing data (No. 9 and 10) while the training data (No. 1-8)

Table 3: Conclusion Table Using Data Testing

1.	Data Testing 9	(W_1.70) (W_2.354) (W_3.3) (W_4.4) (W5.3) (W6.4) (W7.2) (W8.2) (W9.2) (W10.0) (W11.0) + b ≥ + 1
2.	Data Testing 10	(W_1.70) (W_2.354) (W_3.3) (W_4.4) (W5.3) (W6.4) (W7.2) (W8.2) (W9.2) (W10.0) (W11.0) + b ≥ + 1

1.	(W_1.70) (W_2.354) (W_3.3) (W_4.4) (W5.3) (W6.4) (W7.2) (W8.2) (W9.2) (W10.0) (W11.0) + b ≥ + 1 (70. -0,00903) (354. -1,335) (3. 56,571) (4. 49.4) (3. -373,04) (4. 782,382) (2. -560,47) (2. - 63,74) (2. -4,406) (0. -146,541) (0. -3,56) + 1,57 = 4784,65
	(Appropriate, because the prediction result is Positive, while the actual data is satisfied or positive)
2.	(W_1.21) (W_2.1255) (W_3.4) (W_4.1) (W5.1) (W6.1) (W7.1) (W8.4) (W9.86) (W10.75) (W11.0) - b ≥ - 1 (21. -0,00903) (1225. -1,335) (4. 56,571) (1. 49.498,44) (1. 782,382) (1. 782,382) (1. -560,47) (4. - 63,74) (86. -4,406) (75. -146,541) (0. -3,56) - 1,57 = -10766,10
	(Appropriate, because the prediction result is negative, while the actual data is not satisfied or negative)

4.2.2 Random Forest Algorithm Steps

1. Split Dataset based on Attribute Value

Split Dataset used is starting from the first attribute, namely Age. Split Dataset here is meant by sorting the arrivals attribute starting from the smallest to the largest value.

Table 4: Split Dataset Based on Attribute

No	Age	Flight Distance	Gate Location	Inflight wifi service	Inflight entertainment	Online support	Ease of Online booking	On-board service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	Satisfaction
1	7	1241	4	1	1	1	1	3	4	0	0	0
6	15	2138	3	2	0	2	2	3	4	0	0	1
10	21	1255	4	1	1	1	1	4	4	86	75	0
7	37	1701	4	5	1	5	5	3	3	1	0	0
3	46	1863	4	1	1	1	1	4	4	20	22	0
5	46	1863	4	1	1	1	1	4	4	20	22	0
2	47	2464	3	0	2	2	3	4	3	310	305	1
4	47	2464	3	0	2	2	3	4	3	310	305	1
8	60	623	3	3	4	3	1	1	1	0	0	1
9	70	354	3	4	3	4	2	2	2	0	0	1

Table 4 explains: that the Dataset will be sorted in this case the sorted is age, which is sorted from smallest to largest.

2. Calculation of Satisfaction Amount Based on Age Attribute

The next step is to calculate the amount of satisfaction based on the arrivals value which has been sorted as follows:

Table 5: Calculation of the Amount of Satisfaction Based on the Value of Arrivals

Age	Satisfaction		Total
	0 (Tidak Puas)	1 (Puas)	
≤ 7	1	0	1
> 7	4	5	9
≤ 15	1	1	2
> 15	4	4	8
≤ 21	2	1	3
> 21	3	4	7
≤ 37	3	1	4
> 37	2	4	6
≤ 46	5	1	6
> 46	0	4	4
≤ 47	5	3	8
> 47	0	2	2
≤ 60	5	4	9
> 60	0	1	1
≤ 70	5	5	10
> 70	0	0	0

Table 5 Explains: the calculation of the number of satisfactions on the number of age values that have been ordered previously, the calculation of the satisfaction value starts from the smallest to the largest, namely >70

3. Performing Gini Index Calculations

After calculating the amount of satisfaction or categorization, the next step is to calculate the Gini index as follows:

Formulas:

$$Gini\ Index = 1 - \sum_{i=1}^{n} (P_i)^2 \qquad (1)$$

What is meant in the formula is 1 – the frequency of each class (which was categorized earlier).

Table 6: Gini Index Calculations

The following is the calculation of the Gini Index as follows:	
1.	Gini Index (Age) ≤ 7) = 1 - $[(\frac{1}{1})^2 + (\frac{0}{1})^2] = 1 - 1 = 0$
2.	Gini Index (Age > 7) = 1 - $[(\frac{4}{9})^2 + (\frac{5}{9})^2] = 1 - (\frac{41}{81}) = (\frac{40}{81}) = 0{,}493$
3.	Gini Index (Age ≤ 15) = 1 - $[(\frac{1}{2})^2 + (\frac{1}{2})^2] = 1 - (\frac{2}{4}) = (\frac{2}{4}) = 0{,}5$
4.	Gini Index (Age > 15) = 1 - $[(\frac{4}{8})^2 + (\frac{4}{8})^2] = 1 - (\frac{32}{64}) = (\frac{1}{2}) = 0{,}5$
5.	Gini Index (Age ≤ 21) = 1 - $[(\frac{2}{3})^2 + (\frac{1}{3})^2] = 1 - (\frac{5}{9}) = (\frac{4}{9}) = 0{,}444$
6.	Gini Index (Age > 21) = 1 - $[(\frac{3}{7})^2 + (\frac{4}{7})^2] = 1 - (\frac{25}{49}) = (\frac{24}{49}) = 0{,}489$
7.	Gini Index (Age ≤ 37) = 1 - $[(\frac{3}{4})^2 + (\frac{1}{4})^2] = 1 - (\frac{10}{16}) = (\frac{6}{16}) = 0{,}375$
8.	Gini Index (Age > 37) = 1 - $[(\frac{2}{6})^2 + (\frac{4}{6})^2] = 1 - (\frac{20}{36}) = (\frac{16}{36}) = 0{,}444$
9.	Gini Index (Age ≤ 46) = 1 - $[(\frac{5}{6})^2 + (\frac{1}{6})^2] = 1 - (\frac{26}{36}) = (\frac{10}{16}) = 0{,}625$
10.	Gini Index (Age > 46) = 1 - $[(\frac{0}{4})^2 + (\frac{4}{4})^2] = 1 - 1 = 0$
11.	Gini Index (Age ≤ 47) = 1 - $[(\frac{5}{8})^2 + (\frac{3}{8})^2] = 1 - (\frac{34}{64}) = (\frac{30}{64}) = 0{,}468$
12.	Gini Index (Age > 47) = 1 - $[(\frac{0}{2})^2 + (\frac{2}{2})^2] = 1 - 1 = 0$
13.	Gini Index (Age ≤ 60) = 1 - $[(\frac{5}{9})^2 + (\frac{4}{9})^2] = 1 - (\frac{41}{81}) = (\frac{40}{81}) = 0{,}493$
14.	Gini Index (Age > 60) = 1 - $[(\frac{0}{1})^2 + (\frac{1}{1})^2] = 1 - 1 = 0$
15.	Gini Index (Age ≤ 70) = 1 - $[(\frac{5}{10})^2 + (\frac{5}{10})^2] = 1 - (\frac{1}{2}) = (\frac{1}{2}) = 0{,}5$
16.	Gini Index (Age > 70) = 1 - $[(0)^2 + (0)^2] = 1 - 0 = 1$

The Gini Index calculation is calculated based on the results of categorizing the previous classes, namely satisfaction based on the age attribute.

The value (1/1) is obtained from the satisfaction value of 0 and the value of age ≤7. For the numerator 1 is obtained from the number of satisfaction 0 and age 7. And the numerator of 1 is obtained from the number of satisfaction 1 and arrivals 7. While the denominator 1 is the total satisfaction value of 0 and 1. While the value (0/1) is obtained from the satisfaction value in the form of 1 and age 7. This calculation is carried out from age 7 to > 70. To calculate the value of age 15 to > 70, we still use the same method as the age value 7.

4. Performing a Gini Split Calculation

At this stage, I look for the best Gini Split value from the previously formed Gini Index calculation. The following is the Gini Split calculation:

Formulas:

$$Gini\ Split = \sum_{i=1}^{p} \frac{ni}{n} Gini(i) \qquad (2)$$

Table 7: Gini Split Calculation

1.	Gini Split (Age = 7) = $(\frac{1}{10})(0) + (\frac{9}{10})(\frac{40}{81}) = 0{,}444$
2.	Gini Split (Age = 15) = $(\frac{2}{10})(\frac{2}{4}) + (\frac{8}{10})(\frac{1}{2}) = 0{,}5$
3.	Gini Split (Age = 21) = $(\frac{3}{10})(\frac{4}{9}) + (\frac{7}{10})(\frac{24}{49}) = (\frac{2}{15}) + (\frac{12}{35}) = (\frac{50}{105}) = 0{,}476$
4.	Gini Split (Age = 37) = $(\frac{4}{10})(\frac{6}{16}) + (\frac{6}{10})(\frac{16}{36}) = (\frac{3}{20}) + (\frac{4}{15}) = (\frac{25}{60}) = 0{,}416$
5.	Gini Split (Age = 46) = $(\frac{6}{10})(\frac{10}{16}) + (\frac{4}{10})(0) = (\frac{3}{8}) + (0) = 0{,}375$
6.	Gini Split (Age = 47) = $(\frac{8}{10})(\frac{36}{64}) + (\frac{2}{10})(0) = (\frac{9}{20}) + (0) = 0{,}45$
7.	Gini Split (Age = 60) = $(\frac{9}{10})(\frac{40}{81}) + (\frac{1}{10})(0) = (\frac{4}{9}) + (0) = 0{,}444$
8.	Gini Split (Age = 70) = $(\frac{10}{10})(\frac{1}{2}) + (0)(1) = (\frac{1}{2}) + (0) = 0{,}5$

The value (1/10) is obtained from the previous table in the form of numerator 1 and denominator 10. The value of numerator 1 is obtained from total satisfaction of 0 and 1, and the value of denominator 10 is obtained from total satisfaction between ages 7, which is 1 and > 7 which is 9. (9 /10) obtained from the previous table in the form of the numerator 9 and the denominator 10 which is 9 is the total satisfaction of ages > 7 and 10 comes from the total satisfaction between the ages of 7 which is 1 and > 7 which is 9. For the values of 0 and (40/81) are the results Gini index calculation aged 7 and >7. The Gini Split calculation will be carried out continuously until the largest value of the split data is 70.

Table 8: Gini Splitting Index

Age	Satisfaction		Total	Gini Index	Gini Splitting Index
	0 (Tidak Puas)	1 (Puas)			
≤ 7	1	0	1	0	0,444
> 7	4	5	9	0,493	
≤ 15	1	1	2	0,5	0,5
> 15	4	4	8	0,5	
≤ 21	2	1	3	0,444	0,476
> 21	3	4	7	0,489	
≤ 37	3	1	4	0,375	0,416
> 37	2	4	6	0,444	
≤ 46	5	1	6	0,625	**0,375 (Optimal Splitting Point)**
> 46	0	4	4	0	
≤ 47	5	3	8	0,468	0,45
> 47	0	2	2	0	
≤ 60	5	4	9	0,493	0,444
> 60	0	1	1	0	
≤ 70	5	5	10	0,5	0,5
> 70	0	0	0	1	0,55

From the table it can be concluded that the best Gini Splitting Index value is 0.375 because to get the optimal value obtained from the calculation of the smallest Gini Index

5. Create a Decision Tree

After getting the Gini Index results, the next step is to create a Decision Tree based on the Gini Splitting Index results as follows:

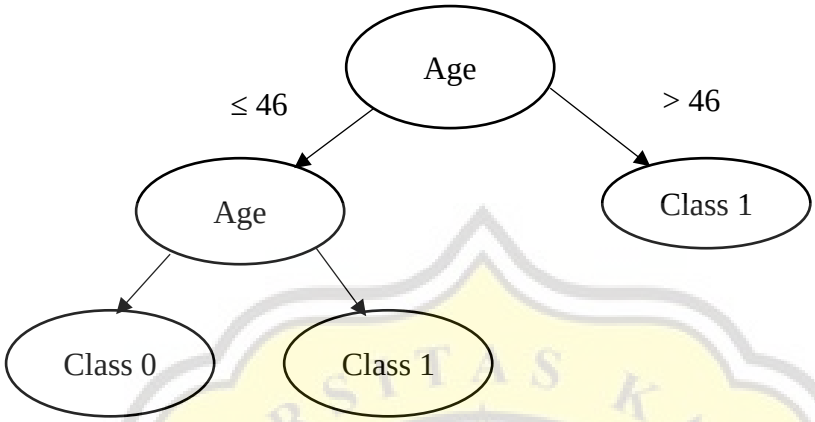

Figure 1: Decision Tree

For the limit values of 46 and > 46, it is determined from the results of the Optimal Gini Split Index, which is 0.375 which comes from age 46 and > 46. The following is a table for age 6 and > 6:

Table 9: Age ≤ 46

No	Age	Flight Distance	Gate Location	Inflight wifi service	Inflight entertainment	Online support	Ease of Online booking	On-board service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	Satisfaction
1	7	1241	4	1	1	1	1	3	4	0	0	0
3	46	1863	4	1	1	1	1	4	4	20	22	0
5	46	1863	4	1	1	1	1	4	4	20	22	0
6	15	2138	3	2	0	2	2	3	4	0	0	1
7	37	1701	4	5	1	5	5	3	3	1	0	0
10	21	1255	4	1	1	1	1	4	4	86	75	0

Table 10: Age > 46

No	Age	Flight Distance	Gate Location	Inflight wifi service	Inflight entertainment	Online support	Ease of Online booking	On-board service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	Satisfaction
2	47	2464	3	0	2	2	3	4	3	310	305	1
4	47	2464	3	0	2	2	3	4	3	310	305	1
8	60	623	3	3	4	3	1	1	1	0	0	1
9	70	354	3	4	3	4	2	2	2	0	0	1

Explanation: Based on the picture of the decision tree and the age table, it can be concluded that grade 1 or age > 46 cannot be detailed because there is only 1 satisfaction among 4 subjects. The solution will occur again if satisfaction is more than 1 as shown in table 5, namely age 46. The decision tree solution will continue to repeat until satisfaction reaches 1 only. To complete satisfaction, it will repeat from the first step of the Random Forest algorithm, which is divided The dataset based on the age attribute is completed until the last step of the decision tree formation and will be repeated also on the second attribute and so on.

6. Random sampling

To find predictions, it is done by making a new sample randomly with the same data, namely 10 data, then repeating from the split step of the dataset whose values are random to the last step forming a decision tree, then comparing the results formed from the original dataset with the new random value dataset

4.3 Accuracy Measurement Support Vector Machine and Random Forest

a. Compute Confusion Matrix

The next step is to calculate the confusion matrix. Confusion Matrix is a performance measurement for machine learning classification problems where the data output is in the form of two or more classes. The confusion matrix is a table that represents 4 different combinations of the predicted value and the actual value. Confusion Matrix describes the predictive value is the output of the program where the value is positive and negative. And the actual value is the actual value where the values are True and False.

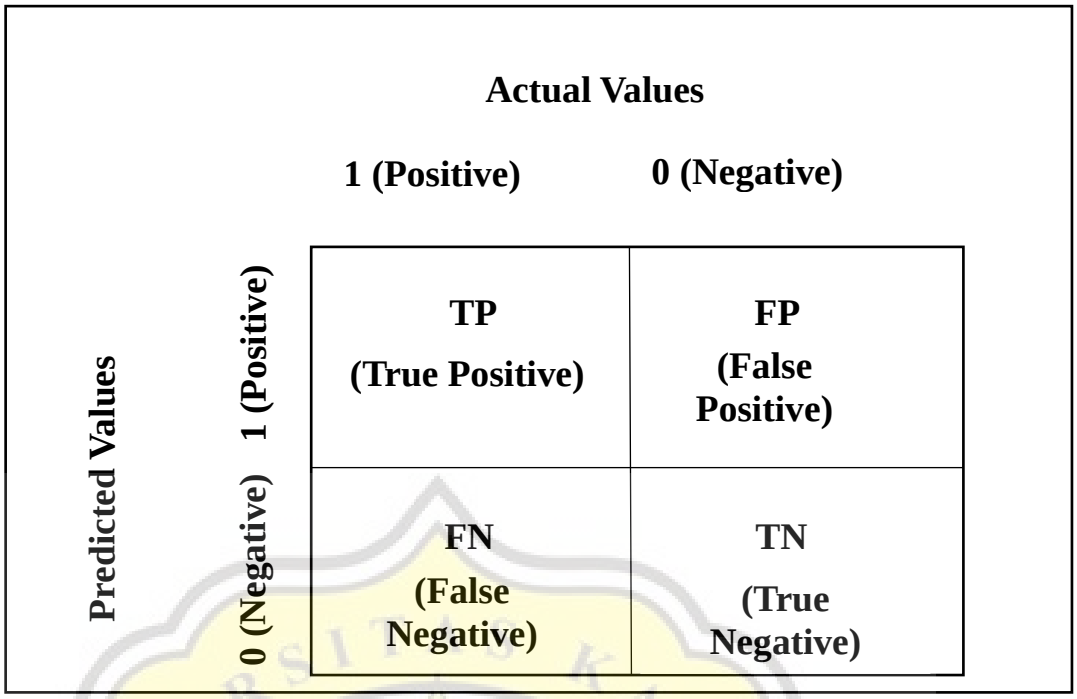

Figure 2: Confusion Matrix

Figure 2 explains the confusion matrix which is divided into 4 parts, namely TP, FP, FN and TN. TP (True Positive) is the actual positive data that is predicted to be correct. FP (False Positive) is the actual positive data that is predicted to be false. FN (False Negative) is the actual negative data that is predicted to be wrong. TN (True Negative) is the actual negative data that is predicted to be true.

From the sample, for example, it produces a confusion matrix as follows:

N=10	Aktual: Positif (1)	Aktual: Negatif (0)
Prediksi Positif: Positif (1)	TP: 7	FP: 1
Prediksi Negatif: Negatif (0)	FN: 1	TN: 1
	8	2

Table 11: Sample Confusion Matrix

b. Generate Accuracy Score, Precision, Recall and F1 Score

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} * 100\% \qquad (3)$$

$$\text{Precision} = \frac{tp}{tp + fp} * 100\%$$

$$\text{Recall} = \frac{tp}{tp + fn} * 100\%$$

$$\text{F1 Score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

In function (6), explaining the value of accuracy is the comparison between data that is classified correctly with all data (how accurate is the model in classifying correctly), the value of precision is the amount of data in the positive category that is classified correctly divided by the total data that is classified as positive (accuracy between the requested data and the prediction results provided by the model), the recall value shows how many percent of the data in the positive category are classified correctly (the model's success in retrieving an information) and the f1 score is a weighted comparison of the average precision and recall . The exact accuracy that is used as a reference for algorithm performance if our dataset has a very close number of False Negative and False Positive data. But if the number is not close, then we should use the F1 Score as a reference.

The following is an example of calculating accuracy, precision, recall and F1-Score:

$$\text{Accuracy: } \frac{7+1}{7+1+1+1} * 100\% = \frac{8}{10} * 100\% = 80\%$$

$$\text{Precision: } \frac{7}{7+1} * 100\% = \frac{7}{8} * 100\% = 87{,}5\%$$

$$\text{Recall: } \frac{7}{7+1} * 100\% = \frac{7}{8} * 100\% = 87{,}5\%$$

$$\text{F1 Score: } \frac{2*0{,}875*0{,}875}{0{,}875+0{,}875} * 100\% = \frac{1{,}53125}{1{,}75} * 100\% = 87{,}5\%$$

c. The Final Result

The final result of this project is to compare the results achieved from the two algorithms in calculating accuracy, precision, recall and f1 score which are the benchmarks for the achievement of this project. If the accuracy value of one of the algorithms reaches a value greater than the other algorithms, then that algorithm is the best algorithm in calculating accuracy, as well as precision, recall and f1 score.

4.2.1. Flowchart SVM Algorithm

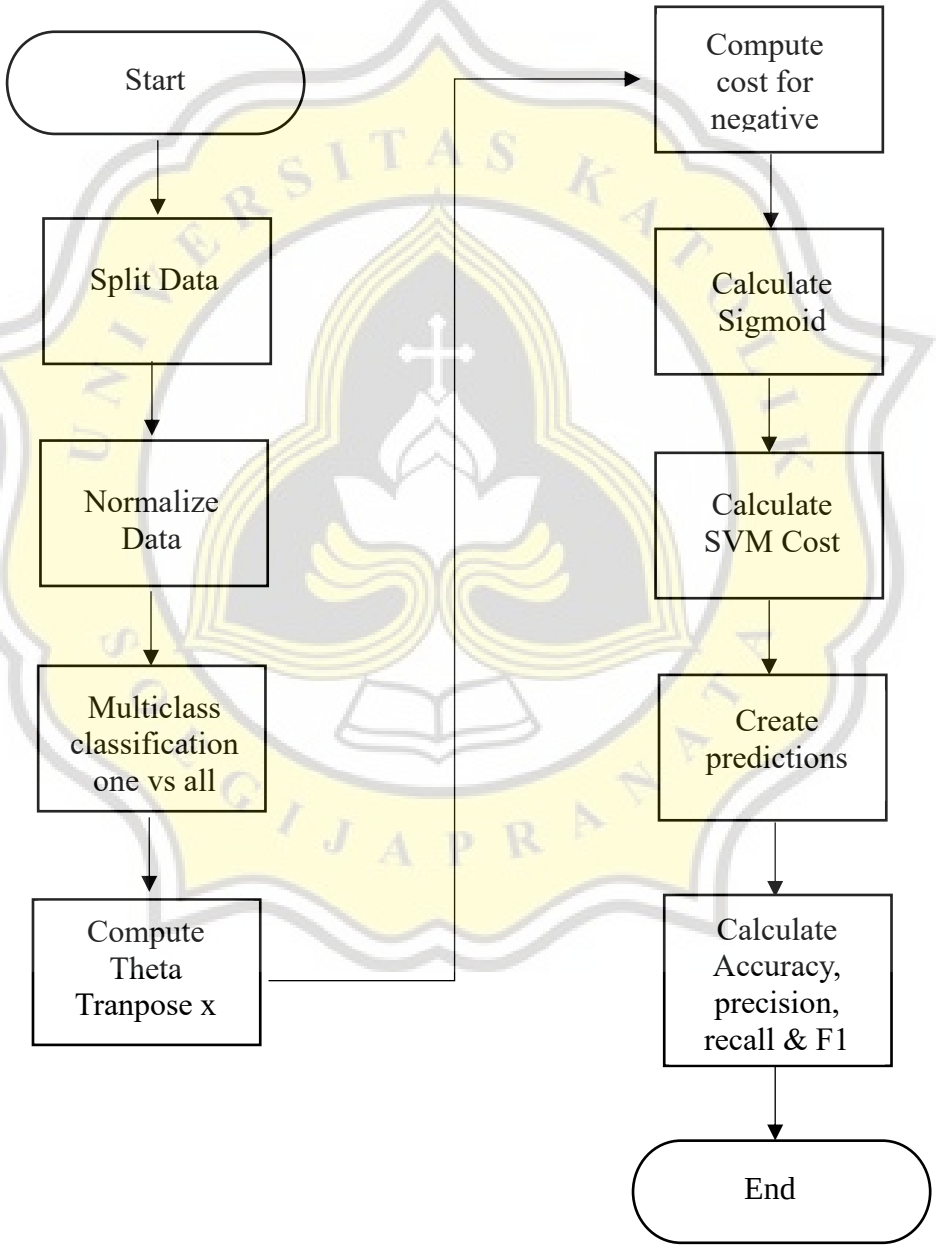

Figure 3: Flowchart SVM Algorithm

In figure 3 above is a flowchart of the svm algorithm starting from data collection. The data is taken from the kaggle website, totaling 11,061 data. The data will be used to apply the SVM algorithm. The stages of applying the algorithm start from the Split Dataset. Split Dataset uses a cross validation split technique which separates the dataset into two parts, namely training and testing data with 70% training data and 30% testing data.

The next step is to use the data normalization technique. The data normalization technique consists of calculating the max-min column for scaling and feature scaling. Min-Max Scaling works by scaling data or adjusting data within a certain range or range. The range that is commonly used is 0 to 1. The next step is to perform multiclass classification with one vs all, in which the number of class labels in the dataset and the number of classifiers generated must be the same.

The next step is to calculate the theta transpose x. Theta transpose x is a statistical function, namely the inverse matrix which inverts the matrix on its diagonal and aims to minimize the cost function. The next step is to calculate the costs for the positive and negative classes. The positive class is marked with a value of 0 and the negative class is marked with a value of 1. This stage will ensure that the margin is achieved up or down. The next step is to calculate the sigmoid value. Sigmoid is an S-shaped curve. Sigmoid will change the z value to be non-linear and have a value from zero to one. The z value is the result of linear regression.

The next step is to calculate the svm cost obtained from theta transpose x, the cost for negative class and positive class, and the sigmoid value. The SVM cost function aims to estimate the logistic function in a piecemeal linear manner. Next do gradient descent. Next is to make predictions based on the gradient descent that is formed. After getting the results of the prediction, the last step is to calculate the accuracy, precision, recall and F1 score based on the confusion matrix formed.

4.2.2. Flowchart Random Forest Algorithm

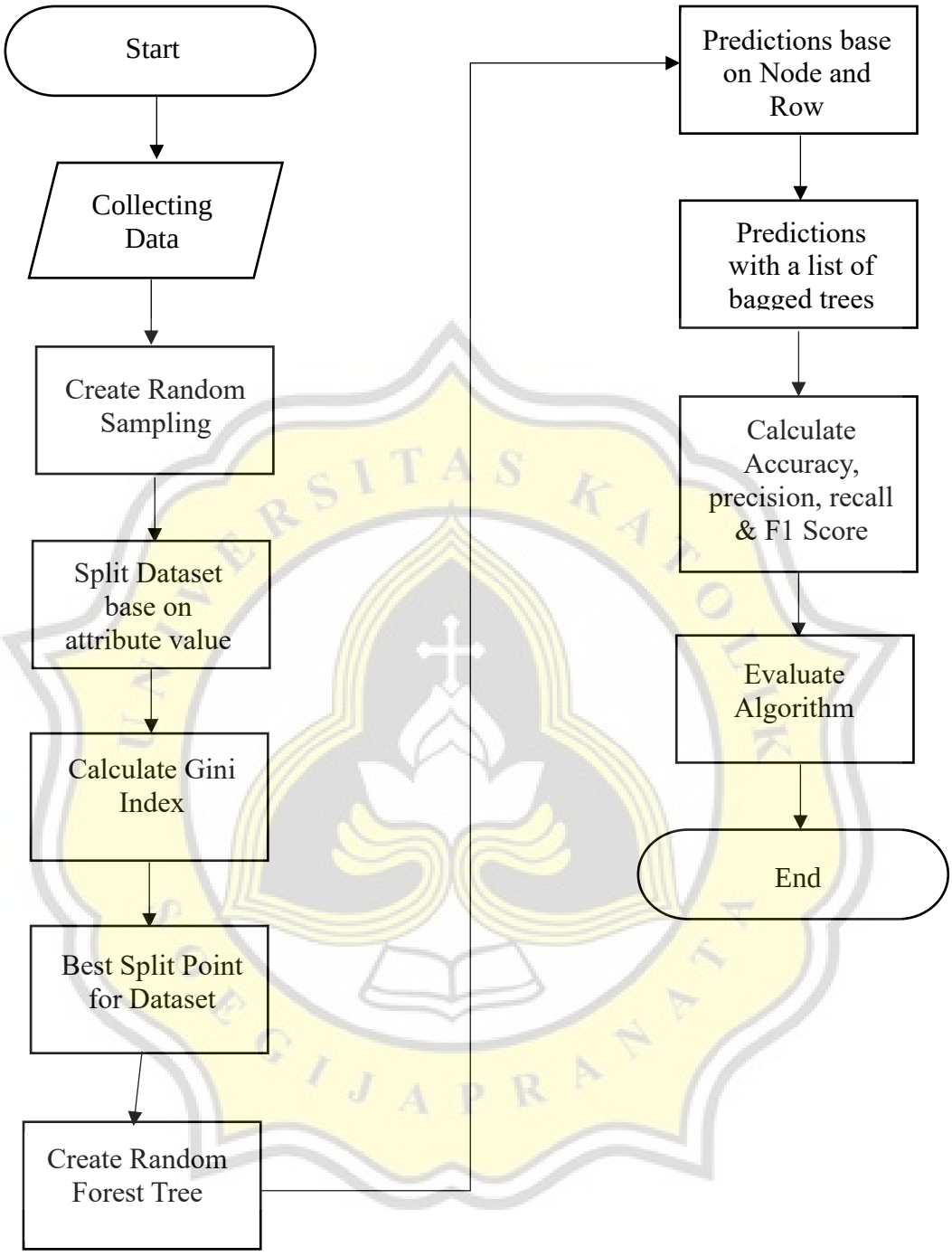

Figure 3:Flowchart Random Forest Algorithm

Random sampling is used for sampling randomly selecting more than one element from a list without repeating elements. It returns a list of unique items selected at random from a list, sequence or set. After creating a random sample, then split the dataset based on the attribute values. Split dataset will separate into two parts, namely left and right.

The next step is to calculate the Gini index value. The Gini index is a measure of variance, the higher the variance, the more misclassification. A lower Gini Index value results in a better classification. The Gini Index measures the degree or probability of certain variables being misclassified when randomly selected. After getting the Gini Index results, the next step is to find the best split point from the dataset. Split point is done repeatedly through the selected features to get the minimum Gini index. The exact Gini index calculation will be sorted from smallest to largest to determine the minimum index Gini value and make split point determination better.

The next step is to create a Random Forest tree, starting with determining the nodes, maximum depth, minimum forest size, root, and terminal nodes. The purpose of making a forest tree is to find out how big the forest is and the depth that is formed. The root node is the top node of the decision tree and the child node or sub node is a division of the root node that is formed because the root node acts as the parent node. The terminal node is at the bottom of the decision tree and has no child nodes.

The next step is to make predictions based on nodes and rows and predictions with bagged tree lists. The difference in predictions between nodes and rows with a bagged tree list is that if a node and row prediction is made, the number of nodes and rows in the index and value will form a prediction in the left or right node, on the other hand predictions with a bagged tree list are predictions formed from the decision tree. In the stored decision tree.

After getting the prediction results, the accuracy, precision, recall and f1 score will be calculated based on the predictions formed. After getting the values of accuracy, precision, recall and f1 score, an evaluation of the model formed will be carried out. To evaluate the model that has been formed, a cross validation split technique is used based on the number of datasets and folds formed. The model evaluation stage also evaluates the prediction results and accuracy to make it better