CHAPTER 5
IMPLEMENTATION AND RESULTS

5.1. Implementation

This experiment was conducted at Google Colab with a time frame of September 19, 2022-October 28, 2022. Before preprocessing the Gaussian Mixture Model, the library needs to be prepared by importing it.

```
1. import numpy as np
2. import pandas as pd
3. from sklearn.mixture import GaussianMixture
4. from sklearn.metrics import classification_report
5. from sklearn.naive_bayes import GaussianNB
6.
7. random = 777
8. np.random.seed(random)
9.
10.  np.set_printoptions(suppress=True)
```

Lines 1-5 import the library used by the author, numpy as numerical computing, and pandas as csv data processing to numerical and vice versa. Lines 7-10 initiate random values in order to get the same results every time you run the program. Line 10 is used to create a fixed numeric value because if it is in exponential form (for example 1e+03) it is considered a string/object so it cannot run properly.

```
11.  loc = "/content/GroundTruthLabelFull.csv"
12.  ref = "/content/ReferencesFull.csv"
13.
14.  # Import CSV to Numpy
15.  data_raw = pd.read_csv(loc).to_numpy()
16.
17.  from sklearn.model_selection import KFold
18.
19.  kf = KFold(n_splits=6, random_state=random, shuffle=True)
20.
21.  train = []
22.  test = []
23.  for train_index, test_index in kf.split(data_raw):
24.      train.append(train_index)
25.      test.append(test_index)
```

Lines 11-15, the dataset in csv format is loaded into the program via pandas and converted to numpy for processing. Next, lines 17-25 are the process of separating into 6 parts for the cross-validator.

```
26. data_ref = pd.read_csv(ref).to_numpy()
27. means, Feature = np.hsplit(data_ref, 2)
```

Lines 26-27 load the reference and separate the absorption band with its polymer as shown in Figure 4.2. The separation is done because the means will be used in the Gaussian Mixture Model as the initial and feature as the list of polymers to look for.

```
28. def processing(p_data):
29.     # Bag of Probabilities / X
30.     bop = []
31.
32.     # Bag of Target
33.     bot = []
34.
35.     # Identification
36.     bodata = bagOfData(p_data)
37.     for dt in bodata:
38.         temp = np.zeros(len(bof))
39.         for x in dt[1:]:
40.             pred = gm.predict([[x]])[0]
41.             ind = bof.index(Feature[pred])
42.             prob = gm.predict_proba([[x]])[0][pred]
43.             if temp[ind] is 0:
44.                 temp[ind] = prob
45.             else:
46.                 if temp[ind] < prob:
47.                     temp[ind] = prob
48.
49.         # probability one hot vector of feature, number label
50.         bot.append(dt[0])
51.         bop.append(temp)
52.
53.     return bop, bot, bodata
```

Lines 28-53 are the function to process the data into a Gaussian Mixture Model that produces a bag of data that is the same size as the existing feature. Lines 40-42 are the process to convert absorption bands into polymers with probability values. Lines 43-47 look for the highest value of a polymer probability, as explained in 4.3.2. Lines 50-53 enter the variable as a separator for each data and return it.

```
54. def microplasticsProba(p_bop):
55.     result = []
56.     for i in range(len(p_bop)):
57.         temp = []
58.         pred = clf.predict_proba(p_bop[i].reshape(1, -1))
59.         max_value = np.amax(pred)
60.         max_index = np.argmax(pred)
61.         temp.append(pred[0])
62.         temp.append(convertClassIdx(max_index))
63.         result.append(temp)
64.
65.     return result
```

Lines 54-65 is a function to convert the polymer opportunity group into microplastics classification with probability. Line 58 is the classification process using Gaussian Naïve Bayes. Line 61-63 is a process to make it easier for readers to see the classification results.

Next for the main process,

```
66. for j in range(len(train)):
67.     data_train = [data_raw[z] for z in train[j]]
68.     data_test = [data_raw[z] for z in test[j]]
69.
70.     data_train = np.array(data_train)
71.     data_test = np.array(data_test)
72.
73.     data = preprocessingTrain(data_train)
```

Line 66 is the function for the cross-validator loop. Lines 67-68 are used to retrieve the data that has been separated from the k-fold index result. Lines 70-71 convert the data into ndarray type, so that the preprocessing function can process the train data and test data.

```
74. bof = []
75.     for fea in Feature:
76.         if fea[0] not in bof:
77.             bof.append(fea[0])
78.
79. x = data_x(data)
```

Lines 74-79 are used to unify polymers of various absorption bands.

```
80.     # Spherical = covariances between its own
81.      gm = GaussianMixture(n_components=means.shape[0], random_state=ran
   dom, means_init=means,covariance_type="spherical")
82.     gm.fit(data)
```

Lines 80-82 is the initiation of the Gaussian Mixture Model, which uses the size of the dataset means and covariance_type spherical. Spherical is used in order to produce a variance value in the covariance.

```
83.     bop, bot, bod = processing(data)
84.
85.     X = bop
86.     Y = bot
87.
88.     clf = GaussianNB()
89.     clf.fit(X, Y)
```

Line 83 is used to obtain the probability gaussian mixture, the target of the dataset, and the whole data. Lines 85-86 separate X and Y and are processed using Gaussian Naïve Bayes on lines 88-89.

```
90. def testing(p_bop, p_bot):
91.     result = []
92.     for i in range(len(p_bop)):
93.         temp = []
94.         pred = clf.predict(p_bop[i].reshape(1, -1))
95.         target = p_bot[i]
96.         temp.append(pred[0])
97.         temp.append(target)
98.         result.append(temp)
99.
100.    return result
```

Lines 90-100 is a function to perform classification and enter it into an array containing targets and predictions. This prediction uses Gaussian Naïve Bayes that has been set on line 88.

```
101.data = preprocessingTrain(data_test)
102.bop, bot, bod = processing(data)
103.result = np.array(testing(bop, bot))
104.print("K-Fold  =  ",  j,  "\n",  classification_report(result[:,1],
   result[:,0]))
```

Lines 101-104 perform the same preprocessing, identification, and classification processes as the train data against the test data. We print the results using the help of sklearn classification_report by entering the target and also the prediction of the result variable on line 104 to produce the classification report.

5.2. Results

From this research, several results were obtained. First, that the center value of a polymer from the Gaussian Mixture Model does not differ much from the existing reference. The range of absorption band values can also be obtained. The results of this range can be used as supporting data for manual matching which can be seen in Table 5.1. The polymer column is the name of the polymer in a microplastics. The reference column is a reference value from previous research, namely Jung et al. [6]. Calculated Means column is the center value of absorption band of a polymer from Gaussian Mixture Model. Calculated Variance column is the variance of the value of the polymer.

Table 5.1. Result of Gaussian Mixture Model

No	Polymer	Reference	Calculated Means	Difference	Calculated Variance
0	C-Cl stretching	700	707.7361	7.7361	18.11433
1	Polar ester groups and benzene ring interaction	712	711.73	0.27	1E-06
2	CH2 rocking	720	721.1141	1.1141	2.381933
3	CH2 rocking	730	729.7961	0.2039	0.864328

4	Aromatic C-H stretching	757	759.6795	2.6795	12.19483
5	Ethyl branching	775	773.7793	1.2207	6.10176
6	Adjacent two aromatic H vibration and aromatic bands	795	794.0669	0.9331	2.662722
7	C-CH3 stretching	840	840.96	0.96	1E-06
8	Aromatic rings 1,2,4,5; Tetra replaced	848	848.8536	0.8536	11.82362
9	Aromatic rings 1,2,4,5; Tetra replaced	872	0	872	0.000001
10	Vinylidene group	890	880.1908	9.8092	59.83227
11	Terminal vinyl group	910	910.0427	0.0427	6.614106
12	CH2 rocking	966	964.41	1.59	1E-06
13	C=C	967	967.3387	0.3387	2.046454
14	C-CH3 rocking	972	974.1955	2.1955	1.673468
15	C-CH3 rocking	997	998.7791	1.7791	0.554114
16	Aromatic C-H bending	1027	1031.516	4.5155	54.2538
17	Methylene group and ester C-O bond vibrations	1050	0	1050	0.000001
18	Methylene group and ester C-O bond vibrations	1096	1099.213	3.2133	3.116134
19	C-C stretching	1099	3.236379	1095.764	3.220798
20	Terephthalate Group (OOCC6H4-COO)	1124	5	1119	1E-06
21	C-CH3 rocking	1165	1169.299	4.2986	0.654167
22	CH2 bending	1199	1194.009	4.9911	0.660411
23	C-O-C	1220	1223.146	3.1457	0.456111
24	Terephthalate Group (OOCC6H4-COO)	1240	1238.3	1.7	1E-06
25	C-H bending	1255	1254.534	0.4658	4.009442
26	C-N stretching	1274	1271.504	2.4964	1.425345
27	C-H bending	1331	1329.915	1.0846	1.788595
28	C-O group stretching of the O-H group deformation and ethylene glycol bending and wagging vibration	1342	1345.948	3.9481	0.567466
29	CH2 bending	1372	1369.003	2.9973	3.395884
30	C-CH3 symmetric	1375	1386.82	11.82	1E-06
31	CH3 groups	1377	1377.723	0.7227	0.909211
32	C-O group stretching of the O-H group deformation and ethylene glycol bending and wagging vibration	1410	1414.34	4.3396	1.30592
33	CH2 scissors	1427	1427.321	0.3211	0.823466

34	CH2 scissors	1435	1434.536	0.4636	1.041862
35	CH2 bending	1451	1452.4	1.4	1E-06
36	C-O group stretching of the O-H group deformation and ethylene glycol bending and wagging vibration	1453	1459.371	6.3710	33.15763
37	CH2 symmetric	1455	1454.33	0.67	0.000001
38	CH2 scissors vibration	1463	1463.97	0.97	9.98E-07
39	CH2 bending	1464	1466.752	2.7518	0.91891
40	CH2 scissors vibration	1475	1475.54	0.54	1.01E-06
41	C=C aromatic stretch	1504	1500.182	3.8180	38.95998
42	N-H bending	1530	1533.916	3.9159	36.45554
43	Aromatic C-H stretching	1547	1540.728	6.2716	2.351811
44	C=C aromatic stretch	1577	1586.251	9.2509	121.0465
45	C=O stretching	1634	1632.578	1.4222	1.849629
46	C=O stretching	1650	1648.816	1.1837	21.4591
47	C=O stretch	1730	1731.115	1.115	8.381026
48	Adjacent two aromatic H vibration and aromatic bands	1960	1959.906	0.0935	4.324141
49	CO2 axial symmetric deformation	2350	2350.907	0.9067	8.367101
50	CH2 symmetric	2838	2840.099	2.0986	7.077286
51	C-H stretching reflects	2850	2850.79	0.79	1.01E-06
52	Symmetric CH2 stretch	2852	2853.42	1.4202	0.889716
53	C-H stretching	2858	2856.546	1.4536	2.162724
54	Symmetric C-H stretch	2908	2910.483	2.4829	1.664519
55	CH2 asymmetric	2917	2922.911	5.9113	1.700227
56	C-H stretching reflects	2923	2922.503	0.4969	2.11109
57	Asymmetric CH2 stretch	2927	2926.01	0.99	1.01E-06
58	C-H stretching	2932	2932.071	0.0712	8.37503
59	CH3 symmetric	2952	2960.43	8.4302	67.7309
60	Symmetric C-H stretch	2969	0	2969	0.000001
61	Symmetric CH stretch	3054	3044.725	9.2747	1905.845
62	Aromatic C-H stretching	3055	3057.607	2.6069	15.43262
63	N-H stretching	3298	3297.725	0.275	9.51709
64	OH group (hydroxyl)	3432	3431.623	0.3768	2.131814

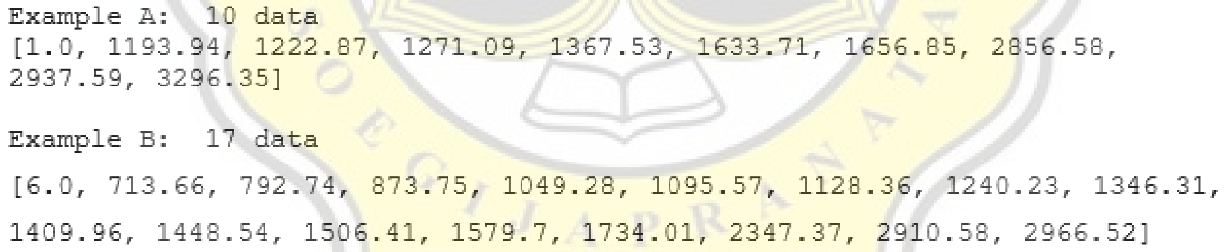
Table 5.1 shows that 59 out of 65 polymers, have a difference in value with the reference of less than 10 with an average of 2.50. Differences of more than 10 and less than 20 is only found in 1 data, namely for polymer C-CH3 symmetric (reference value 1375). In addition, there are 5 polymer data that have a very far distance with the reference, namely, Aromatic rings 1,2,4,5; Tetra

replaced (872), Methylene group and ester C-O bond vibrations (1050), C-C stretching (1099), Terephthalate Group (OOCC6H4-COO) (1124), and Symmetric CH stretch (2969). Therefore, the Gaussian Mixture Model is not much different from the reference.

Too large a deviation can be caused by the absence of data in the vicinity of the polymer from the experimental material. It can be seen from numbers 17 and 60 which have no absorption band value, as well as number 19 with a value of 3.236. In addition to the zero value, outliers can be a factor that makes the absorption band value deviate. Outlier data will make the average value of data in a polymer shift.

Second, variance shows the distribution of data in the polymer. In simple terms, we know what range of absorption band values are included in a polymer. For example, Table 5.1 number 5, the absorption band value for Ethyl branching is in the range of 771.3091 to 776.2495 with the highest probability at 773.7793. This shows agreement with the reference, which is 775. From this result, we can answer the ambiguity of which polymer to classify at manual matching.

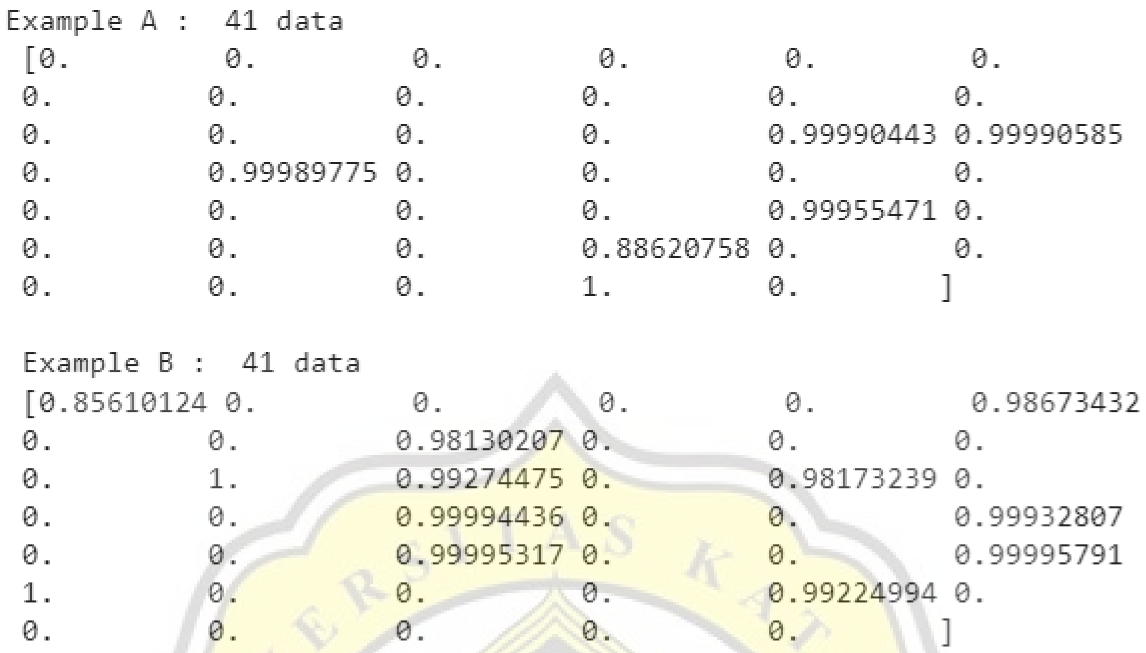
Third, the different data lengths in the dataset (e.g. Figure 5.1) can be equalized through the preprocessing applied. The different data lengths are unified into an array that has a probability value for all polymers as shown in Figure 5.2. The probability value points to the polymers in Table 5.1 which has been grouped by polymer name into Table 5.2.



```
Example A:  10 data
[1.0, 1193.94, 1222.87, 1271.09, 1367.53, 1633.71, 1656.85, 2856.58,
2937.59, 3296.35]

Example B:  17 data
[6.0, 713.66, 792.74, 873.75, 1049.28, 1095.57, 1128.36, 1240.23, 1346.31,
1409.96, 1448.54, 1506.41, 1579.7, 1734.01, 2347.37, 2910.58, 2966.52]
```

Figure 5.1 Example of different size data before preprocessing



```
Example A :  41 data
 [0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.99990443 0.99990585
 0.         0.99989775 0.         0.         0.         0.
 0.         0.         0.         0.         0.99955471 0.
 0.         0.         0.         0.88620758 0.         0.
 0.         0.         0.         1.         0.        ]

 Example B :  41 data
 [0.85610124 0.         0.         0.         0.         0.98673432
 0.         0.         0.98130207 0.         0.         0.
 0.         1.         0.99274475 0.         0.98173239 0.
 0.         0.         0.99994436 0.         0.         0.99932807
 0.         0.         0.99995317 0.         0.         0.99995791
 1.         0.         0.         0.         0.99224994 0.
 0.         0.         0.         0.         0.        ]
```

Figure 5.2 Example of same size data after preprocessing

Table 5.2. Polymer Grouping

No	Polymer
1	C-Cl stretching
2	Polar ester groups and benzene ring interaction
3	CH2 rocking
4	Aromatic C-H stretching
5	Ethyl branching
6	Adjacent two aromatic H vibration and aromatic bands
7	C-CH3 stretching
8	Aromatic rings 1,2,4,5; Tetra replaced
9	Vinylidene group
10	Terminal vinyl group
11	C=C
12	C-CH3 rocking
13	Aromatic C-H bending
14	Methylene group and ester C-O bond vibrations
15	C-C stretching
16	Terephthalate Group (OOCC6H4-COO)
17	CH2 bending
18	C-O-C
19	C-H bending
20	C-N stretching
21	C-O group stretching of the O-H group deformation and ethylene glycol bending and wagging vibration
22	C-CH3 symmetric
23	CH3 groups
24	CH2 scissors
25	CH2 symmetric
26	CH2 scissors vibration
27	C=C aromatic stretch
28	N-H bending
29	C=O stretching
30	C=O stretch
31	CO2 axial symmetric deformation
32	C-H stretching reflects
33	Symmetric CH2 stretch
34	C-H stretching
35	Symmetric C-H stretch
36	CH2 asymmetric
37	Asymmetric CH2 stretch
38	CH3 symmetric
39	Symmetric CH stretch
40	N-H stretching
41	OH group (hydroxyl)

Finally, although the Gaussian Mixture Model has some polymers that are far from the reference, the performance of Gaussian Naïve Bayes obtained from Classification Report by Scikit-learn shows a value of 1.0 which indicates that this model can do its job very well. These results can be seen in 0.

In Table 5.3, K-means is used to compare the performance of Gaussian Mixture because both can be used for identification. However, K-means is a hard clustering which means it has no probability. In addition, Decision Tree is one of the classification methods which in this case is also used as a comparator for Naïve Bayes. To simplify, there are also the graphic of the report on Figure 5.3 and Figure 5.4.

Table 5.3. K-Fold Classification Report

Parameter	***Gaussian Mixture + Naïve Bayes***	***Gaussian Mixture + Decision Tree***	***Kmeans + Naïve Bayes***	***Kmeans + Decision Tree***
Accuracy	1	0.96572	1	0.97619
Precision	1	0.9615	1	0.981996
Recall	1	0.95858	1	0.966534
F1-score	1	0.95806	1	0.976207
Accuracy Sdtev	0	0.03724	0	0.011664
Precision Sdtev	0	0.047546	0	0.005977
Recall Sdtev	0	0.047112	0	0.015035
F1-score Sdtev	0	0.049158	0	0.014254

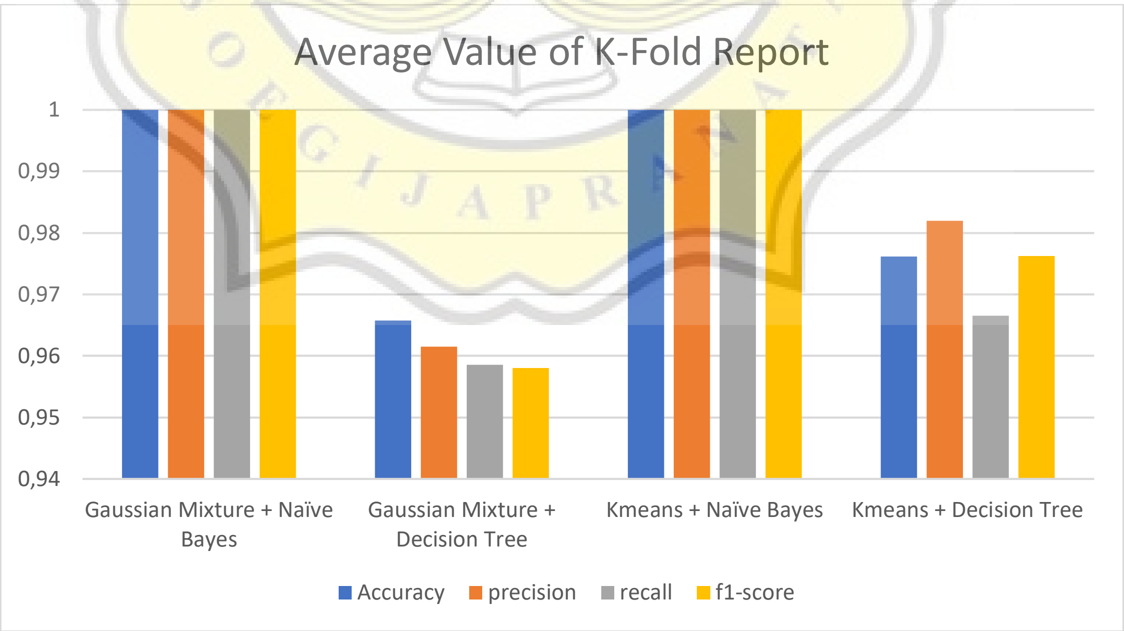


Figure 5.3 Chart of Average Value K-fold report

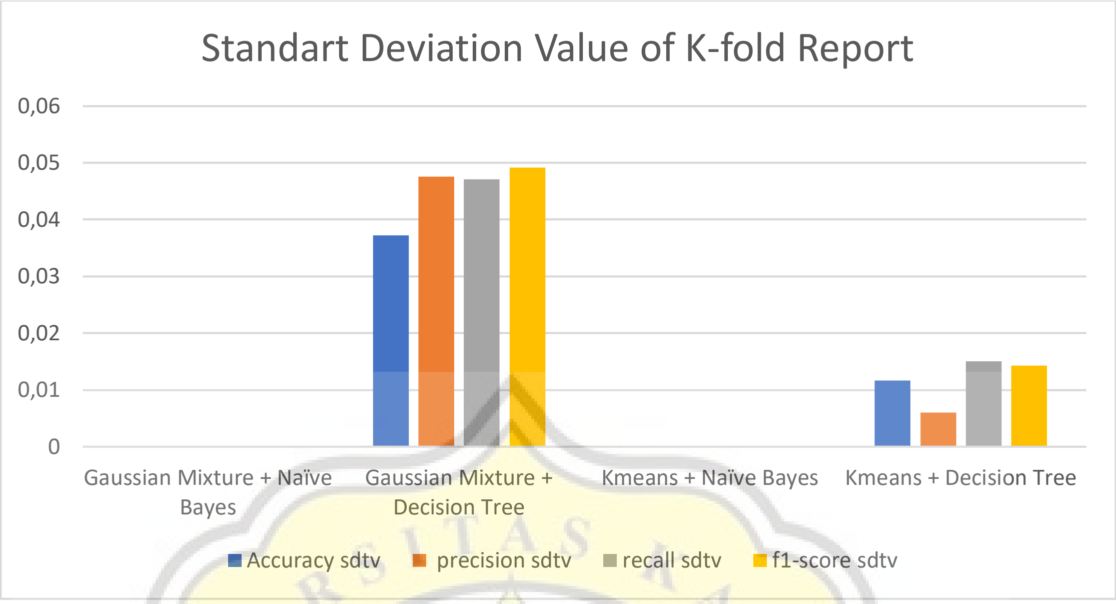


Figure 5.4 Chart of Standart Deviation Value K-fold report

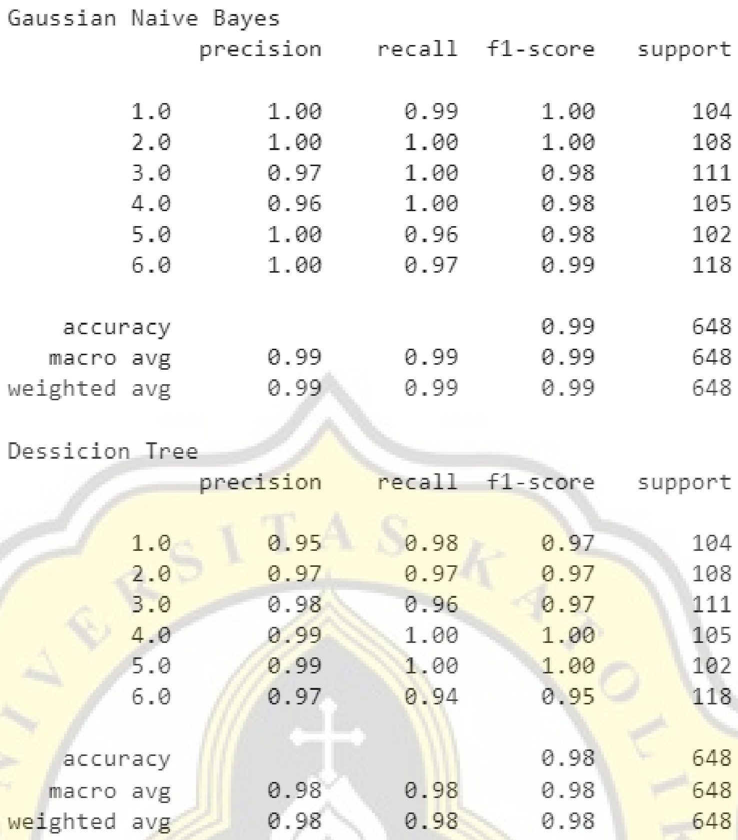
From the comparison, it can be seen that the use of the Gaussian Mixture and K-Means is not much different. If we look at Figure 5.2, we can get the probability value of a polymer up to 100%, while K-means forces the data into one polymer type. This will be a problem if the dataset used is not clean or has a lot of noise. Since the dataset in this case was done in a laboratory with low contamination, this problem does not arise. In the use of Gaussian Mixture, more information is obtained such as, polymer variance and chance value to support the manual matching process so that it is more accurate.

On the other hand, the use of Naïve Bayes is better than Decision Tree because it manages to perform the classification process better even though the values are not much different.

After generate 9 new data for each data for the data augmentation, the accuary of two model did not change differently and it can be seen on below, 0. For the full report of the model on number of augmentation 9 can be seen on Figure 5.5.

Table 5.4. Number of augmentation and the model accuracy

Num of Augmentation	GNB Accuracy	DT Accuracy
2	99	95
3	99	97
4	100	96
5	99	98
6	99	99
7	98	98
8	100	98
9	99	98
10	99	99
11	99	99
12	99	99
13	100	99
14	100	99
15	99	99
16	99	99
18	99	99
20	99	99
23	99	99
25	99	99
30	99	99
50	99	100
100	99	100



```
Gaussian Naive Bayes
              precision    recall  f1-score   support

         1.0       1.00      0.99      1.00       104
         2.0       1.00      1.00      1.00       108
         3.0       0.97      1.00      0.98       111
         4.0       0.96      1.00      0.98       105
         5.0       1.00      0.96      0.98       102
         6.0       1.00      0.97      0.99       118

    accuracy                           0.99       648
   macro avg       0.99      0.99      0.99       648
weighted avg       0.99      0.99      0.99       648

Dessicion Tree
              precision    recall  f1-score   support

         1.0       0.95      0.98      0.97       104
         2.0       0.97      0.97      0.97       108
         3.0       0.98      0.96      0.97       111
         4.0       0.99      1.00      1.00       105
         5.0       0.99      1.00      1.00       102
         6.0       0.97      0.94      0.95       118

    accuracy                           0.98       648
   macro avg       0.98      0.98      0.98       648
weighted avg       0.98      0.98      0.98       648
```

Figure 5.5 Classification Report of Num 9 Augmentation

5.3. Discussion

From the data above, it produces good results because the mean value of the Gaussian Mixture Model is close to the reference [6] by 59 out of 65. In addition, the process of equalizing the component data size is achieved by changing the absorption band data to the highest probability of a polymer. This achievement also makes the performance of Naïve Bayes perfect with an accuracy value of 100% for 6 K-fold.

However, this perfect result can be different if done in different places and times. Because the absorption band value can be significantly different from the existing reference depending on the climate and weather as well as the contamination contained in the sample according to Song et al. [9] in their research.