1. Introduction:

Project Objective: The objective is to build a classification model to determine the accuracy of the classification from a particular dataset based on Naïve Bayes, KNN, and Decision Tree.

Project Description: In Data Mining, Classification is such thing that points out through data mining how can we construct a model based on some predicting attributes. In this project, we have selected the dataset from Kaggle and Completed all related Data Mining processes using Weka. To complete this project, we have taken the dataset “Play Store APK” from Kaggle which analyzes the user satisfaction of a particular app based on its attributes. We have used Weka to preprocess them and classify them using KNN, Naïve Bayes, and Decision tree.

2. Project Outcome:

Possible outcome: For our project, we will implement supervised learning models to our play store mobile app analysis dataset. By using three types of classification models, we might get a proper idea of which classification method would be best for prediction and give a more accurate value. In the end, the comparison between all classification models will give a better understanding.

3. Dataset Details

Name of dataset: *Google Play store Dataset.*

A data set is a collection of numbers or values that relate to a particular subject. In the case of tabular data, a data set corresponds to one or more database tables. In our project, we used real data. Real data is data from a production system, vendor, public records, or any other dataset which otherwise contains operational data. we are working on the public records dataset. From Play Store’s download records and others attributes, we can predict which types of mobile applications are mostly used in our daily life as well as which one is preferable for other users who want to use it. In our project, we consider 7 attributes and 999 instances from the dataset. Rows represent the value of an instance. Seven attributes are -category, Rating, Size, Installs, Types, Content Rating, and Android Version. A special attribute is a sentiment which is the class or target attribute for our project .our dataset is a labeled dataset, where one attribute is given special significance(Sentiment) and the aim is to predict objectives. From dataset,

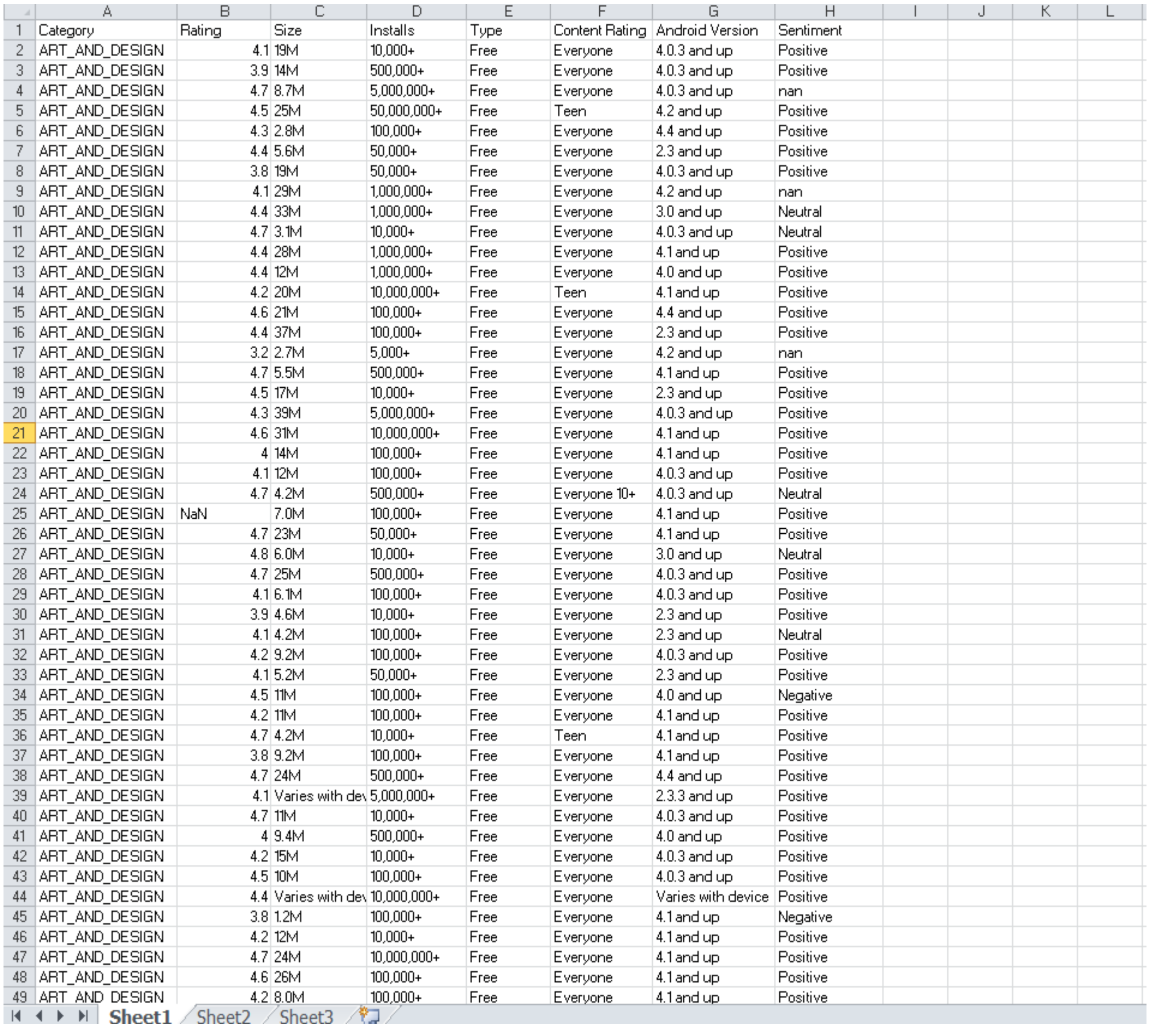
In a nutshell, we have

- Labelled Dataset
- Instance =999
- Attribute = 7
- Target attribute/class = Sentiment

- Attribute (7) = category, Rating, Size, Installs, Types, and Content Rating, Android Version

Dataset Link:

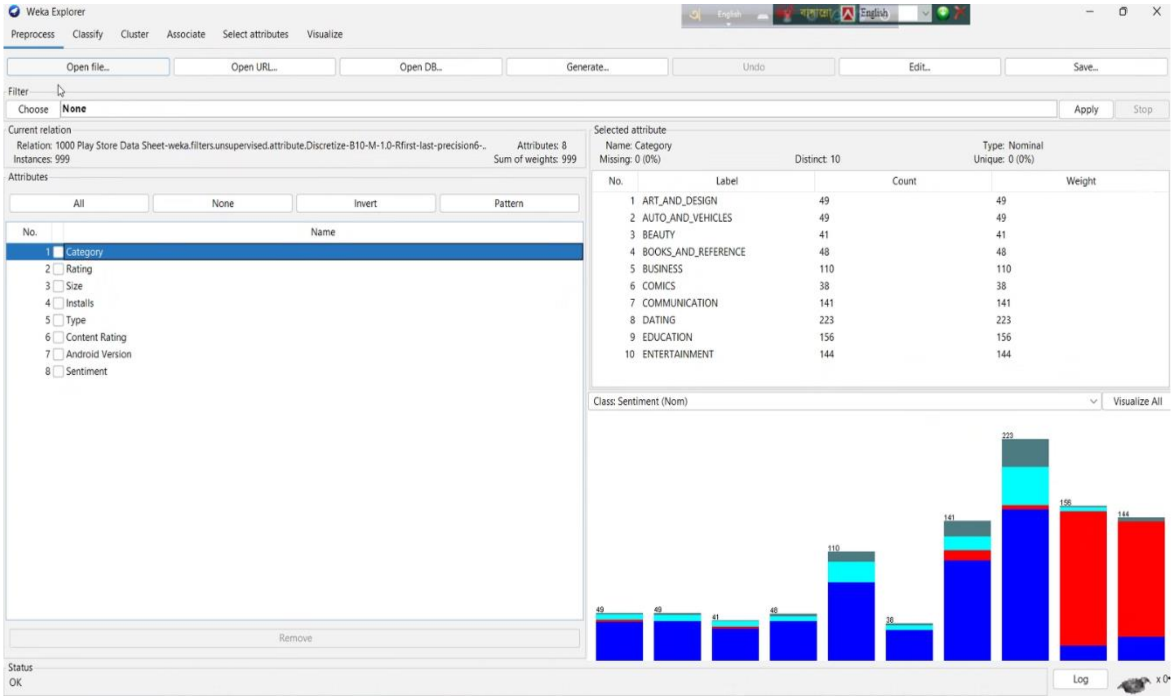
https://www.kaggle.com/datasets/lava18/google-play-store-apps?fbclid=IwAR2PCbqFonzuvFrsbCtTPXK97X1XqTntKfLAra0ADvuooQZcviL_tM-VzOY



	A	B	C	D	E	F	G	H	I	J	K	L
1	Category	Rating	Size	Installs	Type	Content Rating	Android Version	Sentiment				
2	ART_AND_DESIGN	4.1	19M	10,000+	Free	Everyone	4.0.3 and up	Positive				
3	ART_AND_DESIGN	3.9	14M	500,000+	Free	Everyone	4.0.3 and up	Positive				
4	ART_AND_DESIGN	4.7	8.7M	5,000,000+	Free	Everyone	4.0.3 and up	nan				
5	ART_AND_DESIGN	4.5	25M	50,000,000+	Free	Teen	4.2 and up	Positive				
6	ART_AND_DESIGN	4.3	2.8M	100,000+	Free	Everyone	4.4 and up	Positive				
7	ART_AND_DESIGN	4.4	5.6M	50,000+	Free	Everyone	2.3 and up	Positive				
8	ART_AND_DESIGN	3.8	19M	50,000+	Free	Everyone	4.0.3 and up	Positive				
9	ART_AND_DESIGN	4.1	29M	1,000,000+	Free	Everyone	4.2 and up	nan				
10	ART_AND_DESIGN	4.4	33M	1,000,000+	Free	Everyone	3.0 and up	Neutral				
11	ART_AND_DESIGN	4.7	3.1M	10,000+	Free	Everyone	4.0.3 and up	Neutral				
12	ART_AND_DESIGN	4.4	28M	1,000,000+	Free	Everyone	4.1 and up	Positive				
13	ART_AND_DESIGN	4.4	12M	1,000,000+	Free	Everyone	4.0 and up	Positive				
14	ART_AND_DESIGN	4.2	20M	10,000,000+	Free	Teen	4.1 and up	Positive				
15	ART_AND_DESIGN	4.6	21M	100,000+	Free	Everyone	4.4 and up	Positive				
16	ART_AND_DESIGN	4.4	37M	100,000+	Free	Everyone	2.3 and up	Positive				
17	ART_AND_DESIGN	3.2	2.7M	5,000+	Free	Everyone	4.2 and up	nan				
18	ART_AND_DESIGN	4.7	5.5M	500,000+	Free	Everyone	4.1 and up	Positive				
19	ART_AND_DESIGN	4.5	17M	10,000+	Free	Everyone	2.3 and up	Positive				
20	ART_AND_DESIGN	4.3	39M	5,000,000+	Free	Everyone	4.0.3 and up	Positive				
21	ART_AND_DESIGN	4.6	31M	10,000,000+	Free	Everyone	4.1 and up	Positive				
22	ART_AND_DESIGN	4	14M	100,000+	Free	Everyone	4.1 and up	Positive				
23	ART_AND_DESIGN	4.1	12M	100,000+	Free	Everyone	4.0.3 and up	Positive				
24	ART_AND_DESIGN	4.7	4.2M	500,000+	Free	Everyone 10+	4.0.3 and up	Neutral				
25	ART_AND_DESIGN	NaN	7.0M	100,000+	Free	Everyone	4.1 and up	Positive				
26	ART_AND_DESIGN	4.7	23M	50,000+	Free	Everyone	4.1 and up	Positive				
27	ART_AND_DESIGN	4.8	6.0M	10,000+	Free	Everyone	3.0 and up	Neutral				
28	ART_AND_DESIGN	4.7	25M	500,000+	Free	Everyone	4.0.3 and up	Positive				
29	ART_AND_DESIGN	4.1	6.1M	100,000+	Free	Everyone	4.0.3 and up	Positive				
30	ART_AND_DESIGN	3.9	4.6M	10,000+	Free	Everyone	2.3 and up	Positive				
31	ART_AND_DESIGN	4.1	4.2M	100,000+	Free	Everyone	2.3 and up	Neutral				
32	ART_AND_DESIGN	4.2	9.2M	100,000+	Free	Everyone	4.0.3 and up	Positive				
33	ART_AND_DESIGN	4.1	5.2M	50,000+	Free	Everyone	2.3 and up	Positive				
34	ART_AND_DESIGN	4.5	11M	100,000+	Free	Everyone	4.0 and up	Negative				
35	ART_AND_DESIGN	4.2	11M	100,000+	Free	Everyone	4.1 and up	Positive				
36	ART_AND_DESIGN	4.7	4.2M	10,000+	Free	Teen	4.1 and up	Positive				
37	ART_AND_DESIGN	3.8	9.2M	100,000+	Free	Everyone	4.1 and up	Positive				
38	ART_AND_DESIGN	4.7	24M	500,000+	Free	Everyone	4.4 and up	Positive				
39	ART_AND_DESIGN	4.1	Varies with dev	5,000,000+	Free	Everyone	2.3.3 and up	Positive				
40	ART_AND_DESIGN	4.7	11M	10,000+	Free	Everyone	4.0.3 and up	Positive				
41	ART_AND_DESIGN	4	9.4M	500,000+	Free	Everyone	4.0 and up	Positive				
42	ART_AND_DESIGN	4.2	15M	10,000+	Free	Everyone	4.0.3 and up	Positive				
43	ART_AND_DESIGN	4.5	10M	100,000+	Free	Everyone	4.0.3 and up	Positive				
44	ART_AND_DESIGN	4.4	Varies with dev	10,000,000+	Free	Everyone	Varies with device	Positive				
45	ART_AND_DESIGN	3.8	1.2M	100,000+	Free	Everyone	4.1 and up	Negative				
46	ART_AND_DESIGN	4.2	12M	10,000+	Free	Everyone	4.1 and up	Positive				
47	ART_AND_DESIGN	4.7	24M	10,000,000+	Free	Everyone	4.1 and up	Positive				
48	ART_AND_DESIGN	4.6	26M	100,000+	Free	Everyone	4.1 and up	Positive				
49	ART_AND_DESIGN	4.2	8.0M	100,000+	Free	Everyone	4.1 and up	Positive				

Fig: Table of the dataset

Here is, Visualization of attributes from the Weka tool- Separately for all attributes:



Filter: None

Current relation
Relation: 1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-.. Attributes: 8
Instances: 999 Sum of weights: 999

Attributes

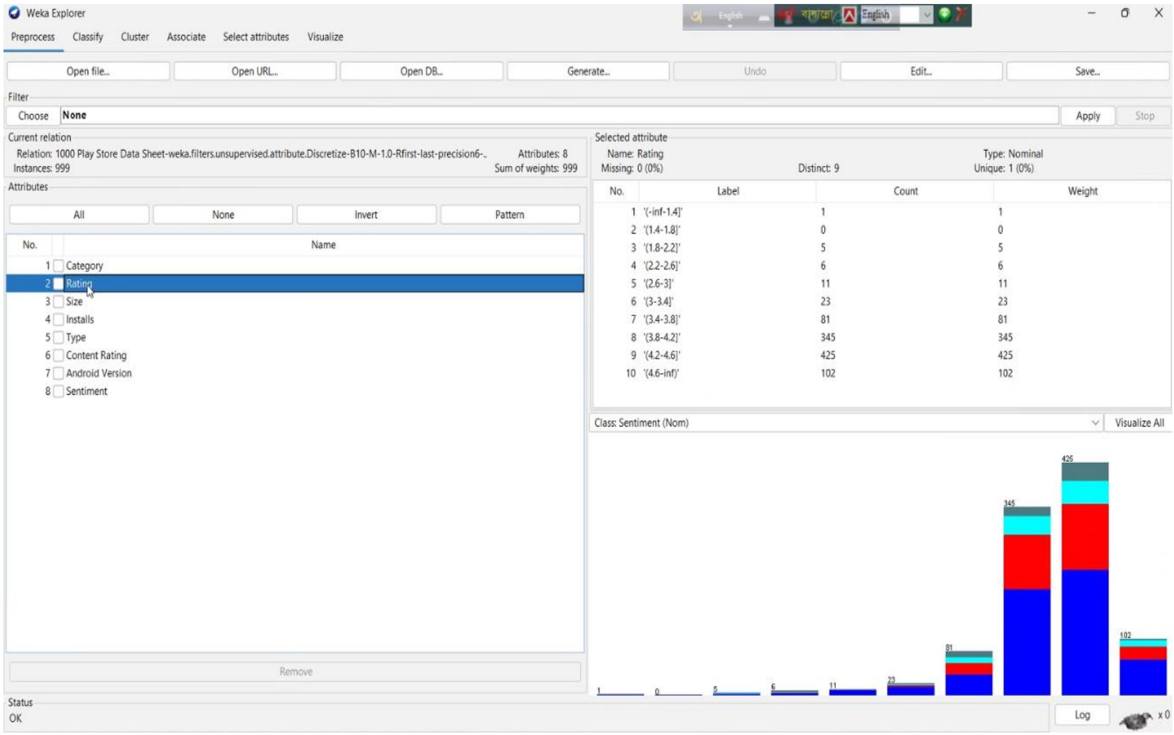
No.	Name
1	☐ Category
2	☐ Rating
3	☐ Size
4	☐ Installs
5	☐ Type
6	☐ Content Rating
7	☐ Android Version
8	☐ Sentiment

Selected attribute
Name: Category Type: Nominal
Missing: 0 (0%) Distinct: 10 Unique: 0 (0%)

No.	Label	Count	Weight
1	ART_AND_DESIGN	49	49
2	AUTO_AND_VEHICLES	49	49
3	BEAUTY	41	41
4	BOOKS_AND_REFERENCE	48	48
5	BUSINESS	110	110
6	COMICS	38	38
7	COMMUNICATION	141	141
8	DATING	223	223
9	EDUCATION	156	156
10	ENTERTAINMENT	144	144

Class: Sentiment (Nom)

Status
OK



Filter: None

Current relation
Relation: 1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-.. Attributes: 8
Instances: 999 Sum of weights: 999

Attributes

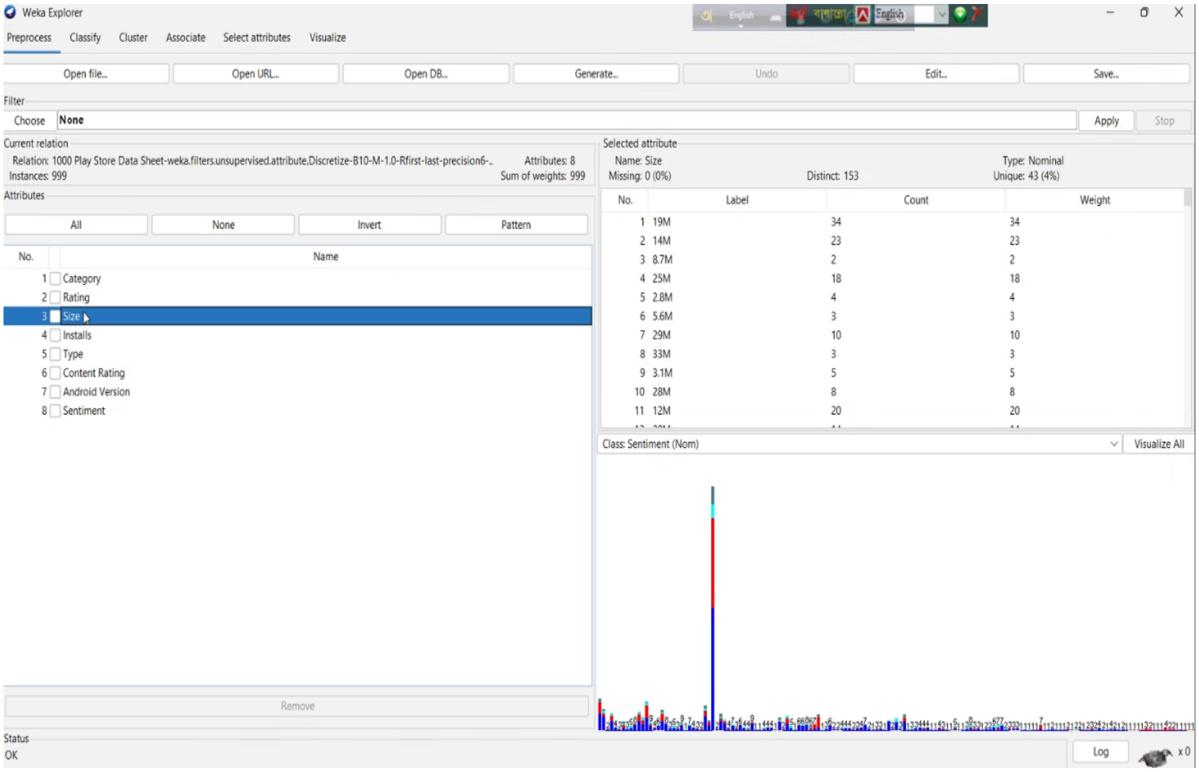
No.	Name
1	☐ Category
2	☐ Rating
3	☐ Size
4	☐ Installs
5	☐ Type
6	☐ Content Rating
7	☐ Android Version
8	☐ Sentiment

Selected attribute
Name: Rating Type: Nominal
Missing: 0 (0%) Distinct: 9 Unique: 1 (0%)

No.	Label	Count	Weight
1	'(-inf-1.4]'	1	1
2	'(1.4-1.8]'	0	0
3	'(1.8-2.2]'	5	5
4	'(2.2-2.6]'	6	6
5	'(2.6-3]'	11	11
6	'(3-3.4]'	23	23
7	'(3.4-3.8]'	81	81
8	'(3.8-4.2]'	345	345
9	'(4.2-4.6]'	425	425
10	'(4.6-inf)'	102	102

Class: Sentiment (Nom)

Status
OK



Filter: None

Current relation

Relation: 1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-.. Attributes: 8
Instances: 999 Sum of weights: 999

Attributes

No.		Name
1	□	Category
2	□	Rating
3	□	Size
4	□	Installs
5	□	Type
6	□	Content Rating
7	□	Android Version
8	□	Sentiment

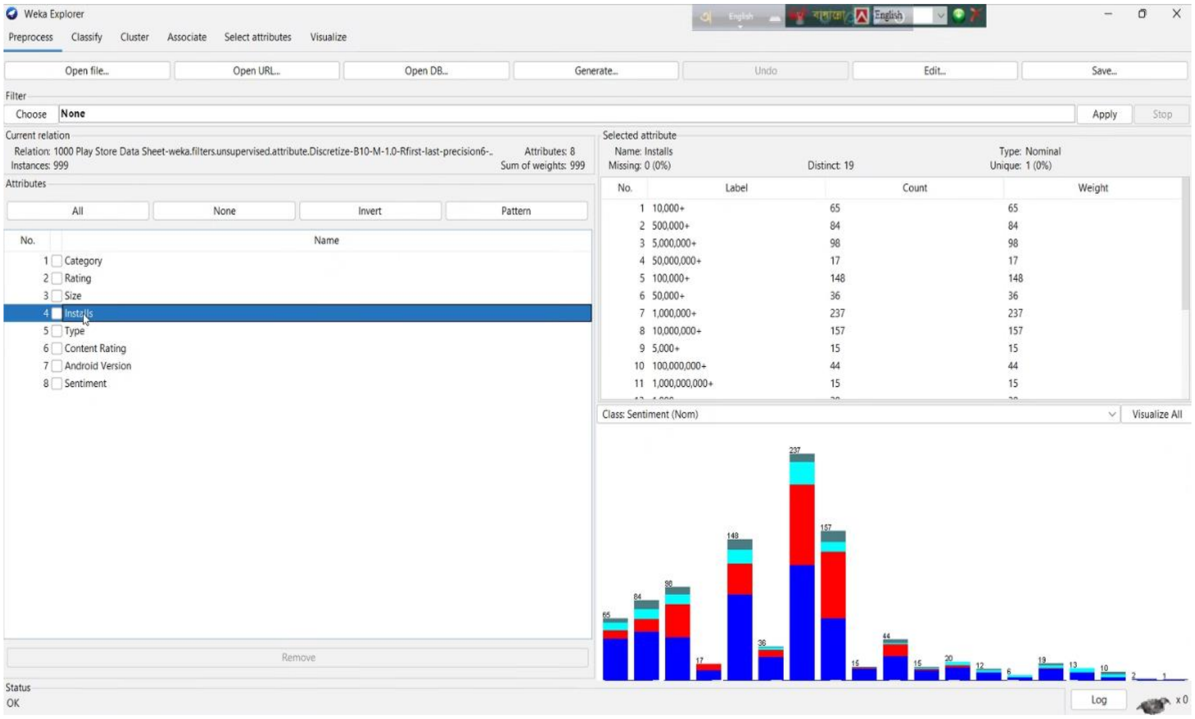
Selected attribute

Name: Size Type: Nominal
Missing: 0 (0%) Distinct: 153 Unique: 43 (4%)

No.	Label	Count	Weight
1	19M	34	34
2	14M	23	23
3	8.7M	2	2
4	25M	18	18
5	2.8M	4	4
6	5.6M	3	3
7	29M	10	10
8	33M	3	3
9	3.1M	5	5
10	28M	8	8
11	12M	20	20

Class: Sentiment (Nom)

Status
OK



Filter: None

Current relation

Relation: 1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-.. Attributes: 8
Instances: 999 Sum of weights: 999

Attributes

No.		Name
1	□	Category
2	□	Rating
3	□	Size
4	□	Installs
5	□	Type
6	□	Content Rating
7	□	Android Version
8	□	Sentiment

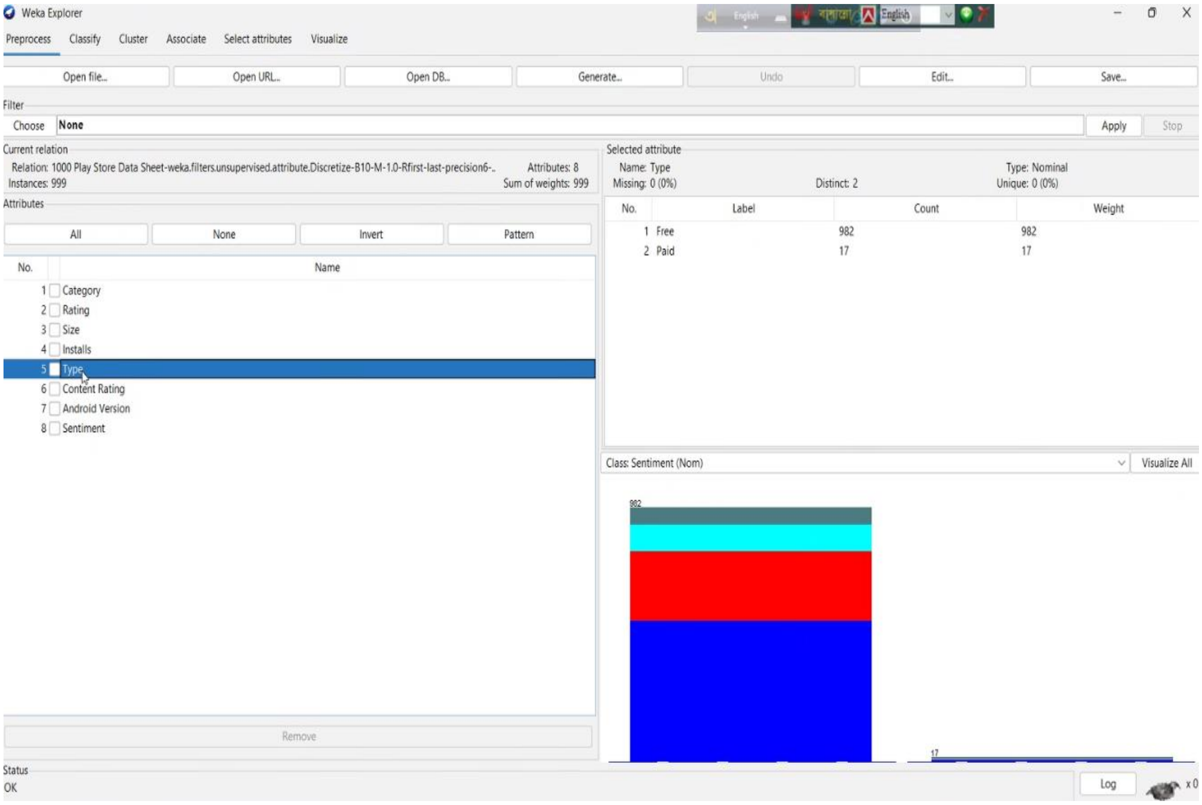
Selected attribute

Name: Installs Type: Nominal
Missing: 0 (0%) Distinct: 19 Unique: 1 (0%)

No.	Label	Count	Weight
1	10,000+	65	65
2	500,000+	84	84
3	5,000,000+	98	98
4	50,000,000+	17	17
5	100,000+	148	148
6	50,000+	36	36
7	1,000,000+	237	237
8	10,000,000+	157	157
9	5,000+	15	15
10	100,000,000+	44	44
11	1,000,000,000+	15	15

Class: Sentiment (Nom)

Status
OK



Filter: None

Current relation
Relation: 1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-... Attributes: 8
Instances: 999 Sum of weights: 999

Attributes

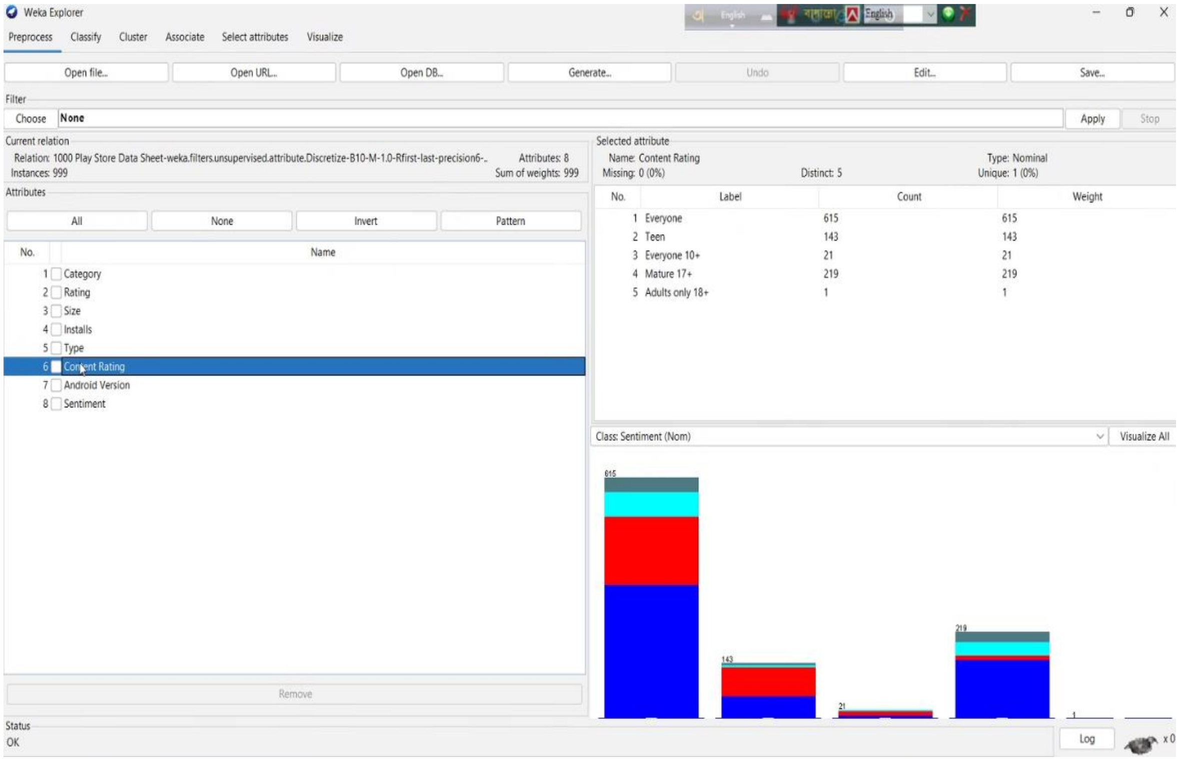
No.	Name
1	Category
2	Rating
3	Size
4	Installs
5	Type
6	Content Rating
7	Android Version
8	Sentiment

Selected attribute
Name: Type Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	Free	982	982
2	Paid	17	17

Class: Sentiment (Nom)

Status
OK



Filter: None

Current relation
Relation: 1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-... Attributes: 8
Instances: 999 Sum of weights: 999

Attributes

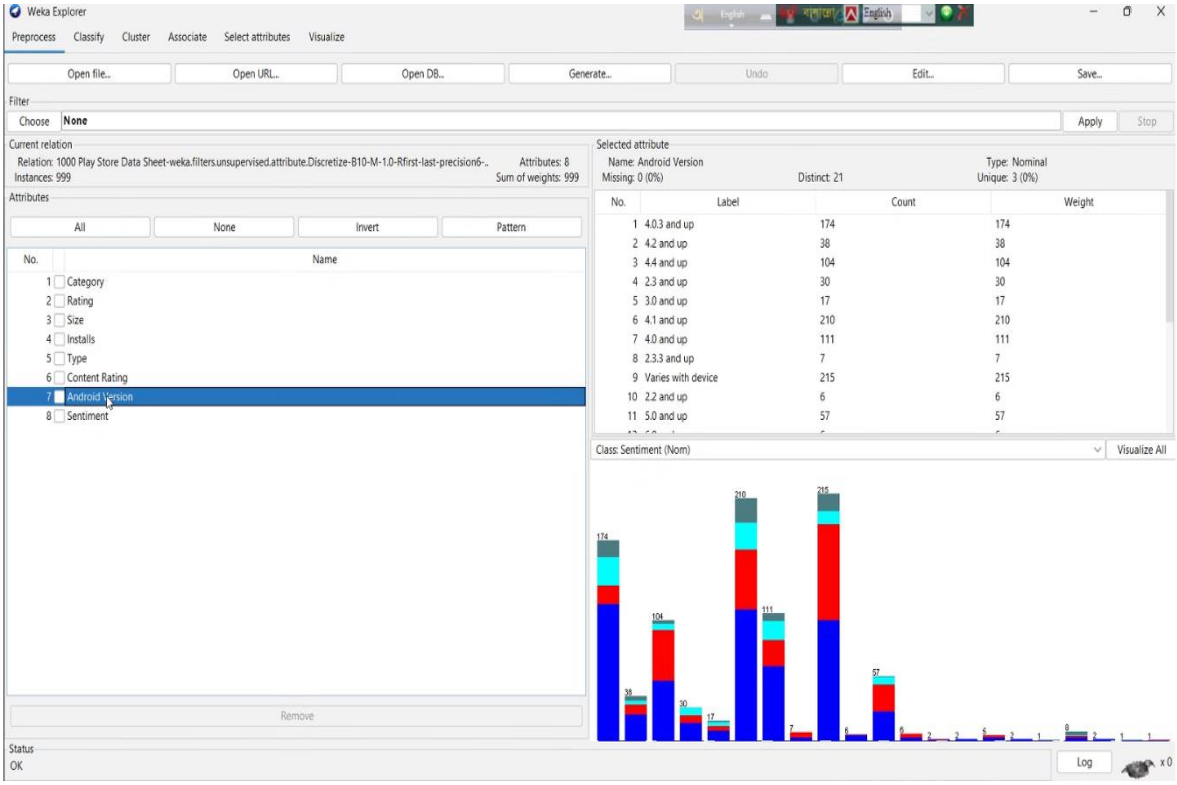
No.	Name
1	Category
2	Rating
3	Size
4	Installs
5	Type
6	Content Rating
7	Android Version
8	Sentiment

Selected attribute
Name: Content Rating Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 1 (0%)

No.	Label	Count	Weight
1	Everyone	615	615
2	Teen	143	143
3	Everyone 10+	21	21
4	Mature 17+	219	219
5	Adults only 18+	1	1

Class: Sentiment (Nom)

Status
OK



Weka Explorer

Current relation

Relation: 1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-.. Attributes: 8
Instances: 999 Sum of weights: 999

Attributes

No.	Name
1	□ Category
2	□ Rating
3	□ Size
4	□ Installs
5	□ Type
6	□ Content Rating
7	□ Android Version
8	□ Sentiment

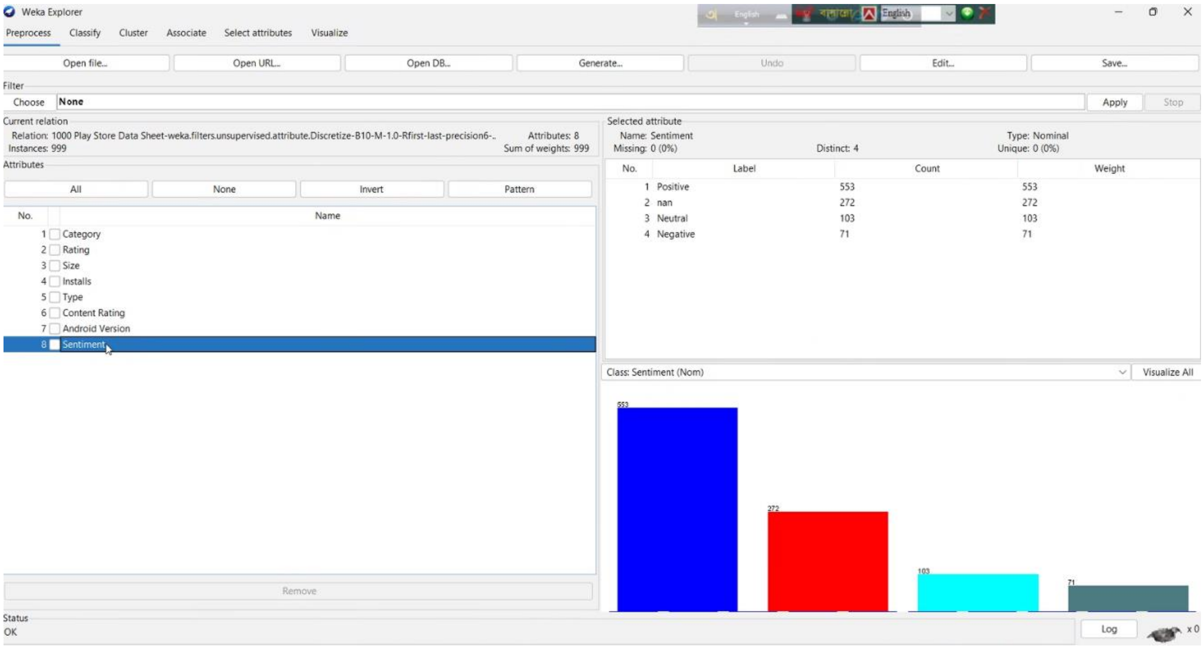
Selected attribute

Name: Android Version Type: Nominal
Missing: 0 (0%) Distinct: 21 Unique: 3 (0%)

No.	Label	Count	Weight
1	4.0.3 and up	174	174
2	4.2 and up	38	38
3	4.4 and up	104	104
4	2.3 and up	30	30
5	3.0 and up	17	17
6	4.1 and up	210	210
7	4.0 and up	111	111
8	2.3.3 and up	7	7
9	Varies with device	215	215
10	2.2 and up	6	6
11	5.0 and up	57	57

Class: Sentiment (Nom)

Status
OK



Weka Explorer

Current relation

Relation: 1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-.. Attributes: 8
Instances: 999 Sum of weights: 999

Attributes

No.	Name
1	□ Category
2	□ Rating
3	□ Size
4	□ Installs
5	□ Type
6	□ Content Rating
7	□ Android Version
8	□ Sentiment

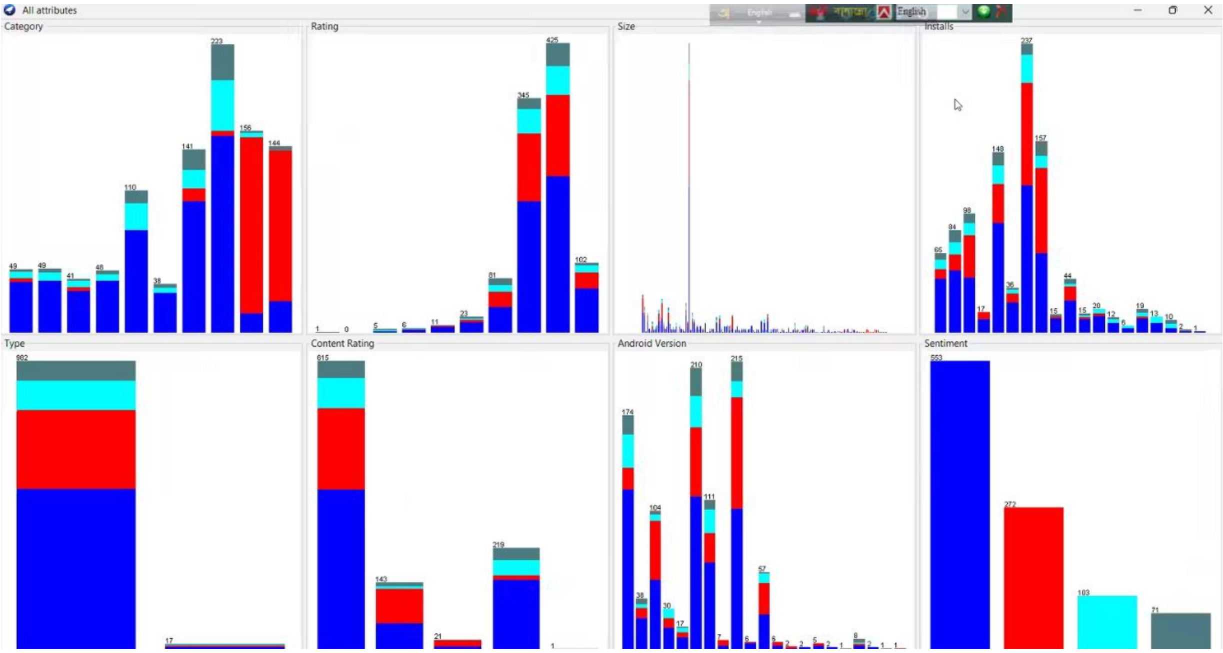
Selected attribute

Name: Sentiment Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count	Weight
1	Positive	553	553
2	nan	272	272
3	Neutral	103	103
4	Negative	71	71

Class: Sentiment (Nom)

Status
OK



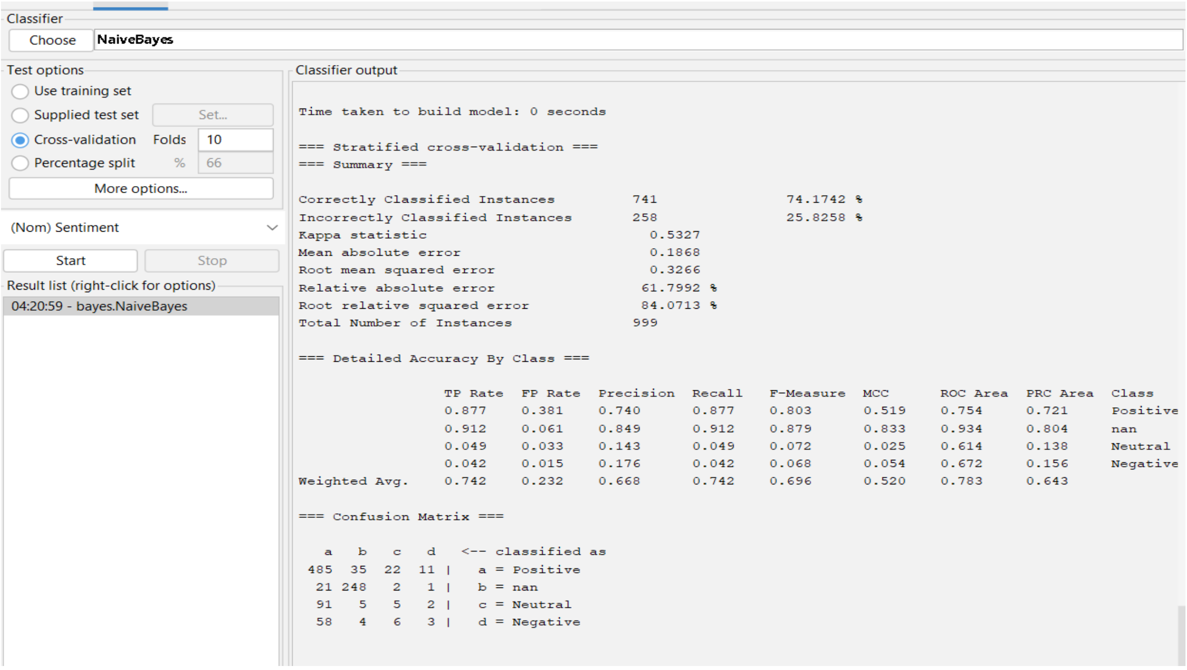
4. Model Development

1. **Data Examine**:
   - ➢ Missing Values:

   In this Dataset, there were a huge number of missing values in the Rating Column. Around 5% of the total data was garbage. In That, case we have replaced it with the most frequent value.

   - ➢ Data Cleaning: In this dataset, there were so many unusual distinct values. We have cleaned the values out.

2. **Classification:**
   Naïve Bayes: The Naïve Bayes classifier constructed based on bays theorem. It’s so easy to build and can easily work on a comparatively large dataset. The classification method of this classifier is more sophisticated that’s why people widely use it. We have performed naïve Bayes classification on our dataset using Weka. And Here’s the result:



Classifier: NaiveBayes

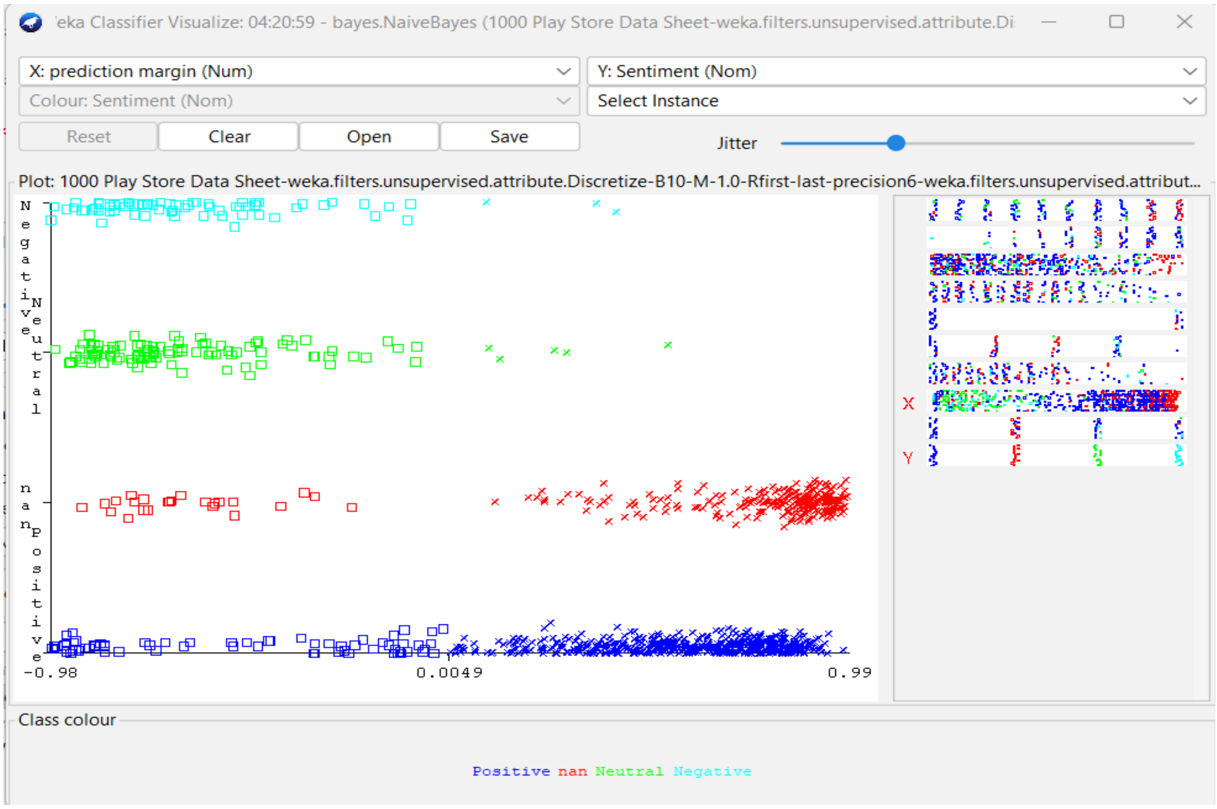
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Sentiment

04:20:59 - bayes.NaiveBayes

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         741               74.1742 %
Incorrectly Classified Instances       258               25.8258 %
Kappa statistic                          0.5327
Mean absolute error                      0.1868
Root mean squared error                  0.3266
Relative absolute error                 61.7992 %
Root relative squared error             84.0713 %
Total Number of Instances              999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.877    0.381    0.740      0.877    0.803      0.519    0.754     0.721     Positive
                 0.912    0.061    0.849      0.912    0.879      0.833    0.934     0.804     nan
                 0.049    0.033    0.143      0.049    0.072      0.025    0.614     0.138     Neutral
                 0.042    0.015    0.176      0.042    0.068      0.054    0.672     0.156     Negative
Weighted Avg.    0.742    0.232    0.668      0.742    0.696      0.520    0.783     0.643

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 485  35  22  11 |   a = Positive
  21 248   2   1 |   b = nan
  91   5   5   2 |   c = Neutral
  58   4   6   3 |   d = Negative
```

Here, we can easily see that, after 10 folds, the technique easily classifies 74% of Data and the mean absolute error is 0.1868.



'eka Classifier Visualize: 04:20:59 - bayes.NaiveBayes (1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Di

X: prediction margin (Num)

Y: Sentiment (Nom)

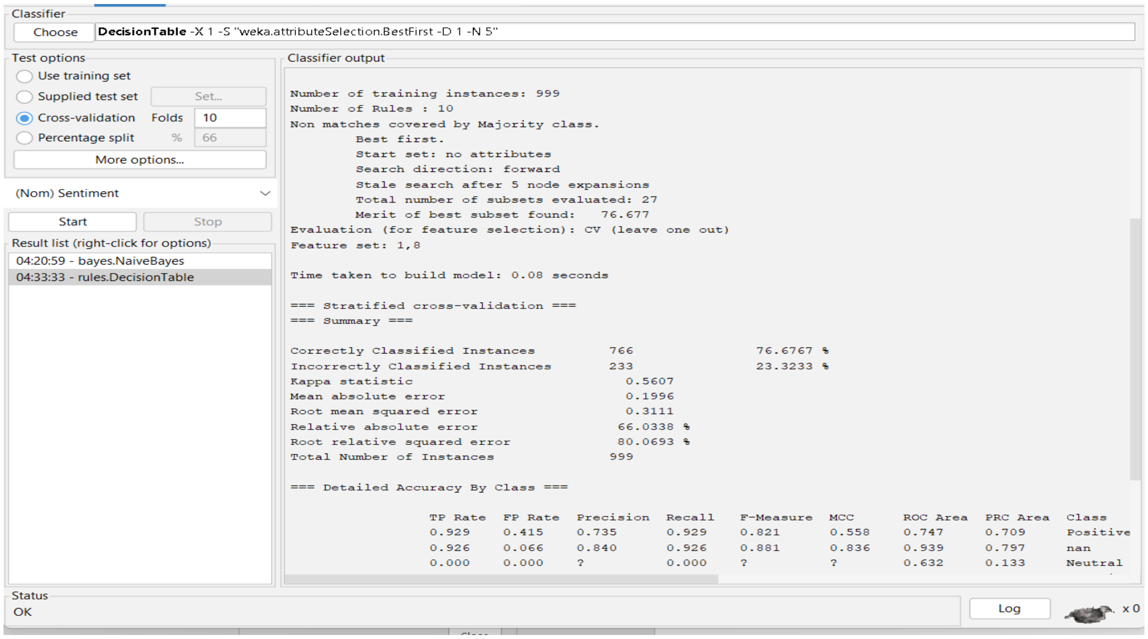
Colour: Sentiment (Nom)

Select Instance

Plot: 1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-weka.filters.unsupervised.attribut...

Herein, we have visualized the classifier error, Naïve Bayes, in the graph. The sentiment is the target variable here and based on that it's showing the error.

Decision Tree: It is a sequential model which always determines the decision that has been made, the probable incident, and the probable outcomes graphically. The major purpose to implement this tree is to determine the best decision for an event. Here we have implemented our dataset through Weka:



Classifier

Choose **DecisionTable** -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"

Test options

○ Use training set
○ Supplied test set Set...
● Cross-validation Folds 10
○ Percentage split % 66

More options...

(Nom) Sentiment

Start Stop

Result list (right-click for options)

04:20:59 - bayes.NaiveBayes
04:33:33 - rules.DecisionTable

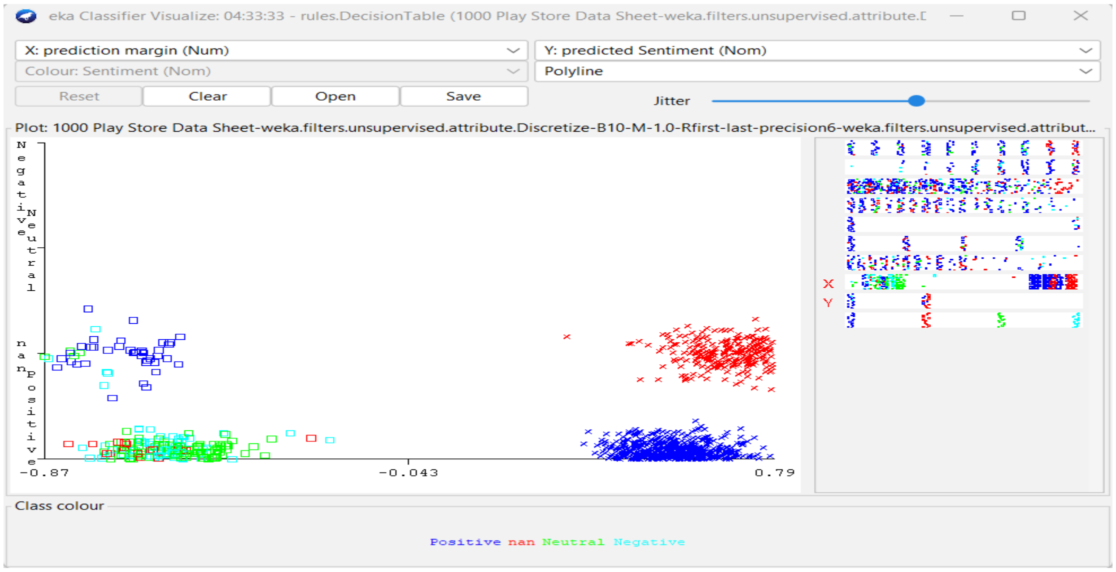
Classifier output

```
Number of training instances: 999
Number of Rules : 10
Non matches covered by Majority class.
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 27
        Merit of best subset found:   76.677
Evaluation (for feature selection): CV (leave one out)
Feature set: 1,8

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         766               76.6767 %
Incorrectly Classified Instances       233               23.3233 %
Kappa statistic                          0.5607
Mean absolute error                      0.1996
Root mean squared error                  0.3111
Relative absolute error                 66.0338 %
Root relative squared error             80.0693 %
Total Number of Instances              999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.929    0.415    0.735      0.929    0.821      0.558    0.747     0.709     Positive
                 0.926    0.066    0.840      0.926    0.881      0.836    0.939     0.797     nan
                 0.000    0.000    ?          0.000    ?          ?        0.632     0.133     Neutral
```

Status

OK

Here, we can easily see that, after 10 folds, the technique easily classifies 76% of the Data and the mean absolute error is 0.1996.



eka Classifier Visualize: 04:33:33 - rules.DecisionTable (1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.[

X: prediction margin (Num)

Y: predicted Sentiment (Nom)

Colour: Sentiment (Nom)

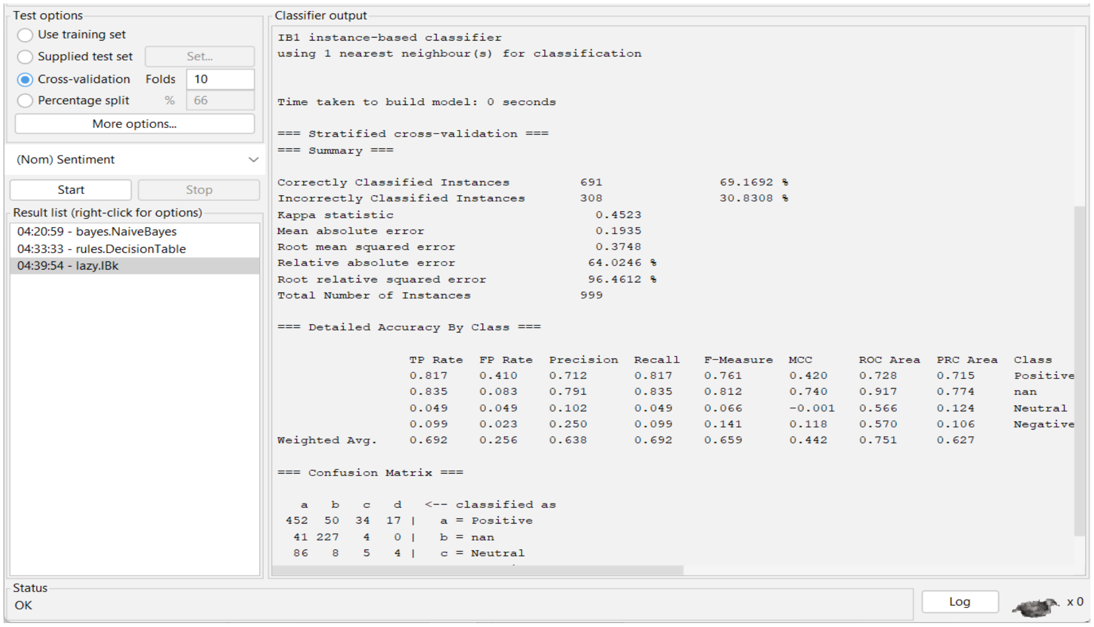
Polyline

Plot: 1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-weka.filters.unsupervised.attribut...

Herein we have visualized the classifier error if the decision tree in the graph. The sentiment is the target variable here and based on that it’s showing the error.

KNN: KNN is a great classification technique. It normally classifies the unseen and new instance based on the previously stored instance using the Euclidian distance measurement technique. Here we have implemented our dataset through Weka:

For 1st nearest Neighbor



Test options
○ Use training set
○ Supplied test set   Set...
● Cross-validation   Folds 10
○ Percentage split   % 66
More options...

(Nom) Sentiment

Start   Stop

Result list (right-click for options)
04:20:59 - bayes.NaiveBayes
04:33:33 - rules.DecisionTable
04:39:54 - lazy.IBk

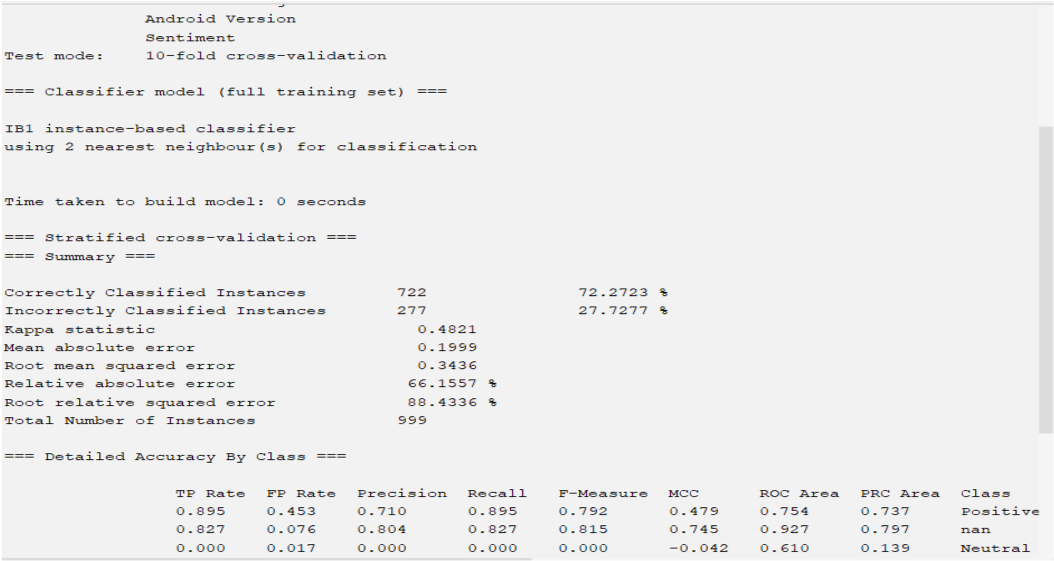
Classifier output

```
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         691               69.1692 %
Incorrectly Classified Instances       308               30.8308 %
Kappa statistic                          0.4523
Mean absolute error                      0.1935
Root mean squared error                  0.3748
Relative absolute error                 64.0246 %
Root relative squared error             96.4612 %
Total Number of Instances              999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.817    0.410    0.712      0.817    0.761      0.420    0.728     0.715     Positive
                 0.835    0.083    0.791      0.835    0.812      0.740    0.917     0.774     nan
                 0.049    0.049    0.102      0.049    0.066      -0.001   0.566     0.124     Neutral
                 0.099    0.023    0.250      0.099    0.141      0.118    0.570     0.106     Negative
Weighted Avg.    0.692    0.256    0.638      0.692    0.659      0.442    0.751     0.627

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 452  50  34  17 |   a = Positive
  41 227   4   0 |   b = nan
  86   8   5   4 |   c = Neutral
```

Status
OK

Here, we can easily see that, after 10 folds, the technique easily classifies 69% of the Data, and the mean absolute error is 0.1935.

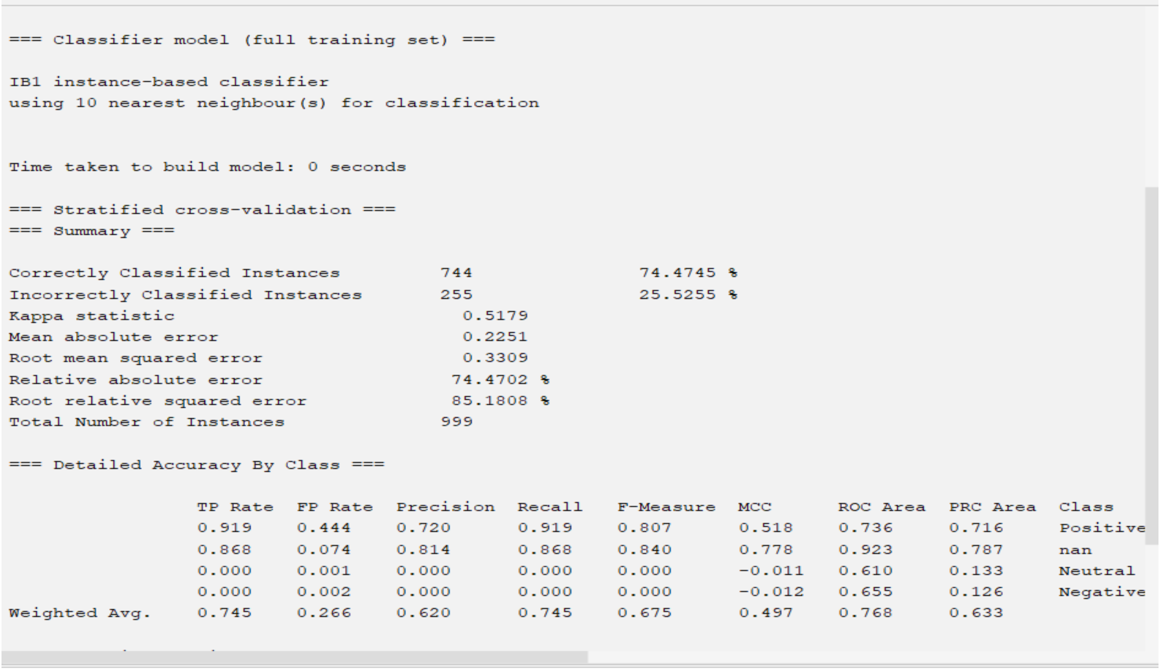
For 2nd Nearest Neighbor



```
              Android Version
              Sentiment
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier
using 2 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         722               72.2723 %
Incorrectly Classified Instances       277               27.7277 %
Kappa statistic                          0.4821
Mean absolute error                      0.1999
Root mean squared error                  0.3436
Relative absolute error                 66.1557 %
Root relative squared error             88.4336 %
Total Number of Instances              999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.895    0.453    0.710      0.895    0.792      0.479    0.754     0.737     Positive
                 0.827    0.076    0.804      0.827    0.815      0.745    0.927     0.797     nan
                 0.000    0.017    0.000      0.000    0.000      -0.042   0.610     0.139     Neutral
```

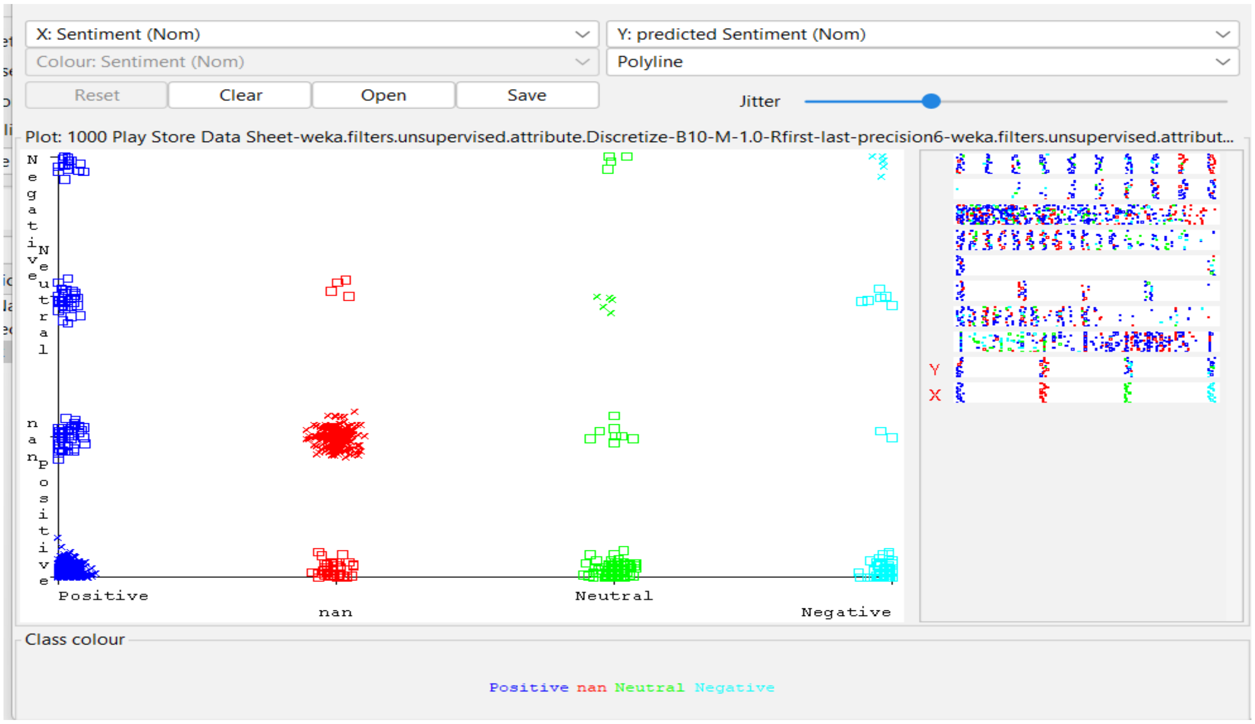
Here, we can easily see that, after 10 folds, the technique easily classifies 72% of the Data, and the mean absolute error is 0.1999.

For 10th Nearest Neighbor



```
=== Classifier model (full training set) ===

IB1 instance-based classifier
using 10 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         744               74.4745 %
Incorrectly Classified Instances       255               25.5255 %
Kappa statistic                          0.5179
Mean absolute error                      0.2251
Root mean squared error                  0.3309
Relative absolute error                 74.4702 %
Root relative squared error             85.1808 %
Total Number of Instances              999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.919    0.444    0.720      0.919    0.807      0.518    0.736     0.716     Positive
                 0.868    0.074    0.814      0.868    0.840      0.778    0.923     0.787     nan
                 0.000    0.001    0.000      0.000    0.000      -0.011   0.610     0.133     Neutral
                 0.000    0.002    0.000      0.000    0.000      -0.012   0.655     0.126     Negative
Weighted Avg.    0.745    0.266    0.620      0.745    0.675      0.497    0.768     0.633
```

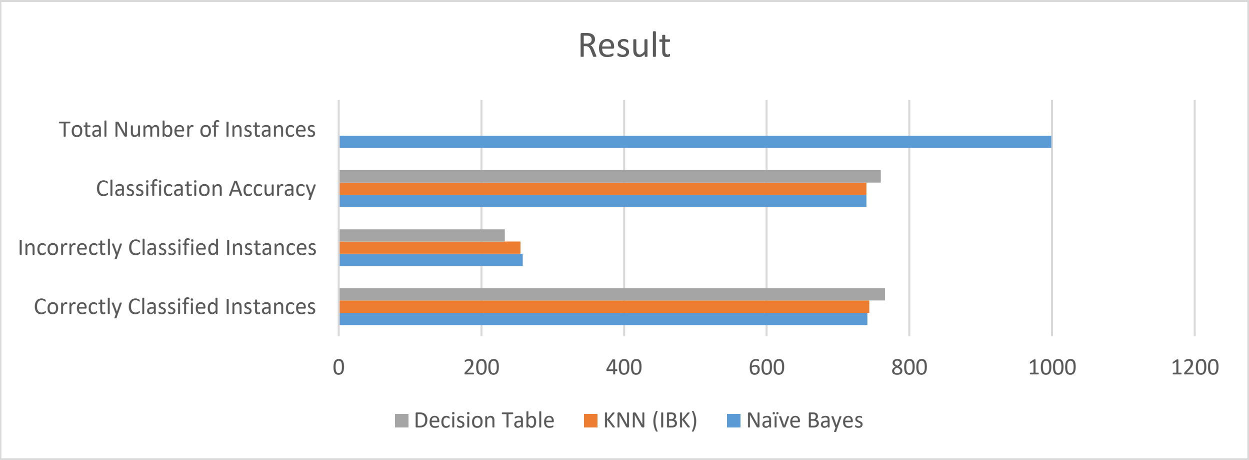
Here, we can easily see that, after 10 folds, the technique easily classifies 74% of the Data, and the mean absolute error is 0.2251.


X: Sentiment (Nom)
Y: predicted Sentiment (Nom)
Colour: Sentiment (Nom)
Polyline
Reset
Clear
Open
Save
Jitter
Plot: 1000 Play Store Data Sheet-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6-weka.filters.unsupervised.attribut...
Negative
Neutral
nan
Positive
Positive
nan
Neutral
Negative
Y
X
Class colour
Positive nan Neutral Negative

Herein we have visualized the classifier error 10 Nearest Neighbor in the graph. The sentiment is the target variable here and based on that it's showing the error.

Result:

Techniques	Folds	Correction	Mean Absolute Error
Naïve Bayes	10	74%	0.1868
Decision Tree	10	76%	0.1996
KNN	10	74%	0.2251



The Below tables and the graph show the accuracy and the error details of the three classification techniques we have used here. Here we can easily see in terms of correction, The correction of Naïve Bayes and KNN(k=10) are the same, but the decision tree gives the best result. Again, in terms of mean absolute error, Naïve Bayes Gives less error.

5. Discussion

The chosen data set has a huge number of missing values in the rating column which were replaced by the most frequent values. And also has garbage values that were replaced in the same manner as the missing value. Unusual values are also cleaned. Naïve Byes, KNN, and Decision Tree classification technique is applied in the dataset in the same settings (10 folds) using Weka and extracted the desired values and then compared with each other to choose the best technique. Naïve Byes produced the best result as it has a 74% accuracy rate and a Mean absolute error of 0.1868). Decision Tree (76% accuracy rate and Mean absolute error 0.1996) produced a better result than KNN but not good as Naïve Byes. KNN produced the worst result (74% accuracy rate and Mean absolute error of 0.2251).

6. Conclusion

Classification fills a very human need to impose order on nature and find hidden relationships. By grouping organisms and species together it was originally hoped that huge masses of data could be stored and retrieved more easily. The accuracy is also an important thing in classification to learn or acknowledge from the classification. For this purpose, Naïve Byes, KNN, and Decision Tree are used in the project to determine which gives the best accuracy in a certain setting. The outcomes show that the accuracy of the Decision tree classification technique is higher (76%) than KNN and Naïve Byes. But Decision tree technique also has the highest mean absolute error of the other two techniques. Overall Naïve Byes produces the best result having a 74% accuracy rate and the least mean absolute error (0.1868).