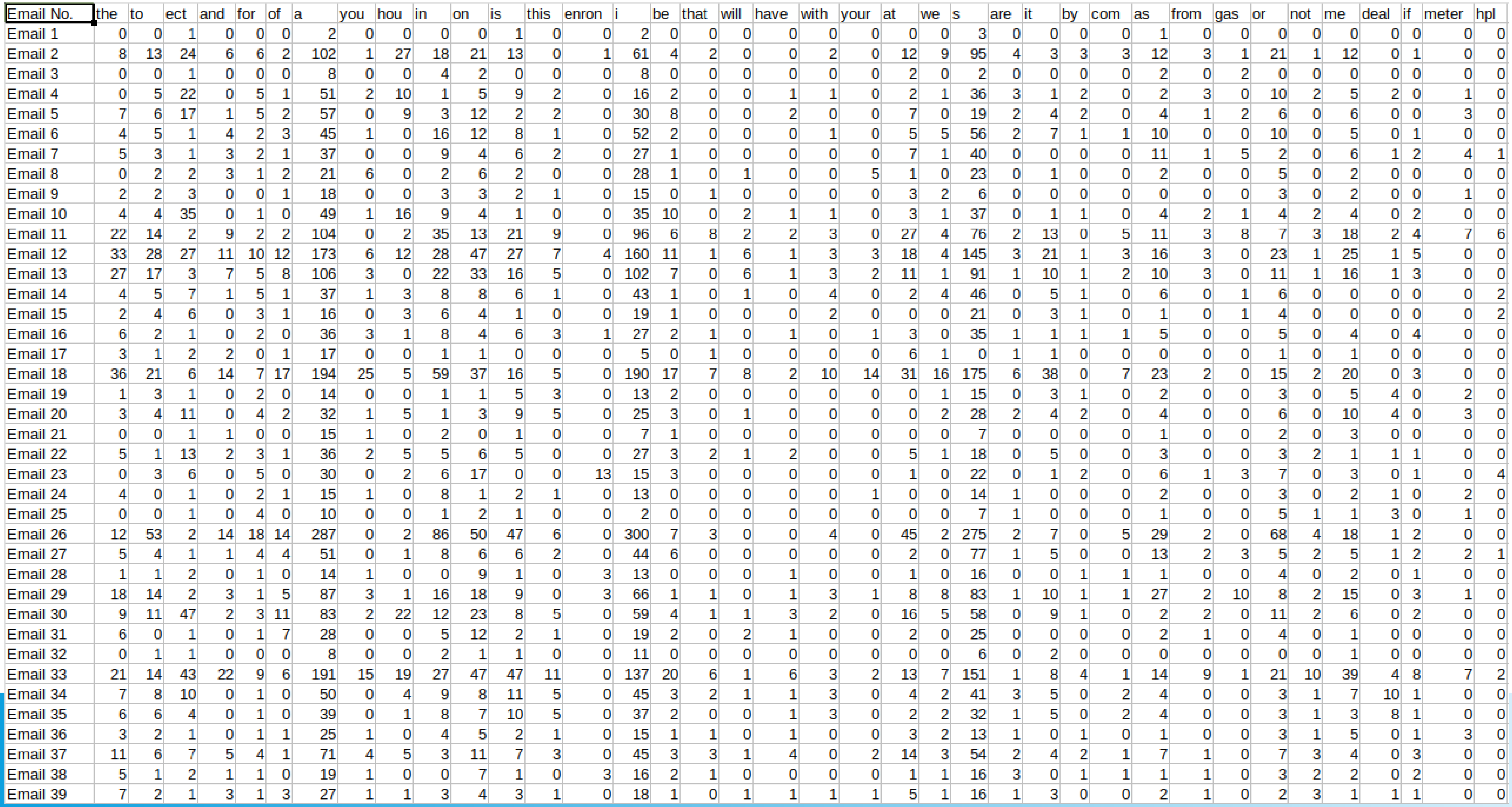
Email Spam Classification

Ashish V Nair

DATASET

- https://www.kaggle.com/balaka18/email-spam-classification-dataset-csv?select=emails.csv



Email No.	the	to	ect	and	for	of	a	you	hou	in	on	is	this	enron	i	be	that	will	have	with	your	at	we	s	are	it	by	com	as	from	gas	or	not	me	deal	if	meter	hpl
Email 1	0	0	1	0	0	0	2	0	0	0	0	1	0	0	2	0	0	0	0	0	0	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Email 2	8	13	24	6	6	2	102	1	27	18	21	13	0	1	61	4	2	0	0	2	0	12	9	95	4	3	3	3	12	3	1	21	1	12	0	1	0	0
Email 3	0	0	1	0	0	0	8	0	0	4	2	0	0	0	8	0	0	0	0	0	0	2	0	2	0	0	0	0	2	0	2	0	0	0	0	0	0	0
Email 4	0	5	22	0	5	1	51	2	10	1	5	9	2	0	16	2	0	0	1	1	0	2	1	36	3	1	2	0	2	3	0	10	2	5	2	0	1	0
Email 5	7	6	17	1	5	2	57	0	9	3	12	2	2	0	30	8	0	0	2	0	0	7	0	19	2	4	2	0	4	1	2	6	0	6	0	0	3	0
Email 6	4	5	1	4	2	3	45	1	0	16	12	8	1	0	52	2	0	0	0	1	0	5	5	56	2	7	1	1	10	0	0	10	0	5	0	1	0	0
Email 7	5	3	1	3	2	1	37	0	0	9	4	6	2	0	27	1	0	0	0	0	0	7	1	40	0	0	0	0	11	1	5	2	0	6	1	2	4	1
Email 8	0	2	2	3	1	2	21	6	0	2	6	2	0	0	28	1	0	1	0	0	5	1	0	23	0	1	0	0	2	0	0	5	0	2	0	0	0	0
Email 9	2	2	3	0	0	1	18	0	0	3	3	2	1	0	15	0	1	0	0	0	0	3	2	6	0	0	0	0	0	0	0	3	0	2	0	0	1	0
Email 10	4	4	35	0	1	0	49	1	16	9	4	1	0	0	35	10	0	2	1	1	0	3	1	37	0	1	1	0	4	2	1	4	2	4	0	2	0	0
Email 11	22	14	2	9	2	2	104	0	2	35	13	21	9	0	96	6	8	2	2	3	0	27	4	76	2	13	0	5	11	3	8	7	3	18	2	4	7	6
Email 12	33	28	27	11	10	12	173	6	12	28	47	27	7	4	160	11	1	6	1	3	3	18	4	145	3	21	1	3	16	3	0	23	1	25	1	5	0	0
Email 13	27	17	3	7	5	8	106	3	0	22	33	16	5	0	102	7	0	6	1	3	2	11	1	91	1	10	1	2	10	3	0	11	1	16	1	3	0	0
Email 14	4	5	7	1	5	1	37	1	3	8	8	6	1	0	43	1	0	1	0	4	0	2	4	46	0	5	1	0	6	0	1	6	0	0	0	0	0	2
Email 15	2	4	6	0	3	1	16	0	3	6	4	1	0	0	19	1	0	0	0	2	0	0	0	21	0	3	1	0	1	0	1	4	0	0	0	0	0	2
Email 16	6	2	1	0	2	0	36	3	1	8	4	6	3	1	27	2	1	0	1	0	1	3	0	35	1	1	1	1	5	0	0	5	0	4	0	4	0	0
Email 17	3	1	2	2	0	1	17	0	0	1	1	0	0	0	5	0	1	0	0	0	0	6	1	0	1	1	0	0	0	0	0	1	0	1	0	0	0	0
Email 18	36	21	6	14	7	17	194	25	5	59	37	16	5	0	190	17	7	8	2	10	14	31	16	175	6	38	0	7	23	2	0	15	2	20	0	3	0	0
Email 19	1	3	1	0	2	0	14	0	0	1	1	5	3	0	13	2	0	0	0	0	0	0	1	15	0	3	1	0	2	0	0	3	0	5	4	0	2	0
Email 20	3	4	11	0	4	2	32	1	5	1	3	9	5	0	25	3	0	1	0	0	0	0	2	28	2	4	2	0	4	0	0	6	0	10	4	0	3	0
Email 21	0	0	1	1	0	0	15	1	0	2	0	1	0	0	7	1	0	0	0	0	0	0	0	7	0	0	0	0	1	0	0	2	0	3	0	0	0	0
Email 22	5	1	13	2	3	1	36	2	5	5	6	5	0	0	27	3	2	1	2	0	0	5	1	18	0	5	0	0	3	0	0	3	2	1	1	1	0	0
Email 23	0	3	6	0	5	0	30	0	2	6	17	0	0	13	15	3	0	0	0	0	0	1	0	22	0	1	2	0	6	1	3	7	0	3	0	1	0	4
Email 24	4	0	1	0	2	1	15	1	0	8	1	2	1	0	13	0	0	0	0	0	1	0	0	14	1	0	0	0	2	0	0	3	0	2	1	0	2	0
Email 25	0	0	1	0	4	0	10	0	0	1	2	1	0	0	2	0	0	0	0	0	0	0	0	7	1	0	0	0	1	0	0	5	1	1	3	0	1	0
Email 26	12	53	2	14	18	14	287	0	2	86	50	47	6	0	300	7	3	0	0	4	0	45	2	275	2	7	0	5	29	2	0	68	4	18	1	2	0	0
Email 27	5	4	1	1	4	4	51	0	1	8	6	6	2	0	44	6	0	0	0	0	0	2	0	77	1	5	0	0	13	2	3	5	2	5	1	2	2	1
Email 28	1	1	2	0	1	0	14	1	0	0	9	1	0	3	13	0	0	0	1	0	0	1	0	16	0	0	1	1	1	0	0	4	0	2	0	1	0	0
Email 29	18	14	2	3	1	5	87	3	1	16	18	9	0	3	66	1	1	0	1	3	1	8	8	83	1	10	1	1	27	2	10	8	2	15	0	3	1	0
Email 30	9	11	47	2	3	11	83	2	22	12	23	8	5	0	59	4	1	1	3	2	0	16	5	58	0	9	1	0	2	2	0	11	2	6	0	2	0	0
Email 31	6	0	1	0	1	7	28	0	0	5	12	2	1	0	19	2	0	2	1	0	0	2	0	25	0	0	0	0	2	1	0	4	0	1	0	0	0	0
Email 32	0	1	1	0	0	0	8	0	0	2	1	1	0	0	11	0	0	0	0	0	0	0	0	6	0	2	0	0	0	0	0	0	0	1	0	0	0	0
Email 33	21	14	43	22	9	6	191	15	19	27	47	47	11	0	137	20	6	1	6	3	2	13	7	151	1	8	4	1	14	9	1	21	10	39	4	8	7	2
Email 34	7	8	10	0	1	0	50	0	4	9	8	11	5	0	45	3	2	1	1	3	0	4	2	41	3	5	0	2	4	0	0	3	1	7	10	1	0	0
Email 35	6	6	4	0	1	0	39	0	1	8	7	10	5	0	37	2	0	0	1	3	0	2	2	32	1	5	0	2	4	0	0	3	1	3	8	1	0	0
Email 36	3	2	1	0	1	1	25	1	0	4	5	2	1	0	15	1	1	0	1	0	0	3	2	13	1	0	1	0	1	0	0	3	1	5	0	1	3	0
Email 37	11	6	7	5	4	1	71	4	5	3	11	7	3	0	45	3	3	1	4	0	2	14	3	54	2	4	2	1	7	1	0	7	3	4	0	3	0	0
Email 38	5	1	2	1	1	0	19	1	0	0	7	1	0	3	16	2	1	0	0	0	0	1	1	16	3	0	1	1	1	1	0	3	2	2	0	2	0	0
Email 39	7	2	1	3	1	3	27	1	1	3	4	3	1	0	18	1	0	1	1	1	1	5	1	16	1	3	0	0	2	1	0	2	3	1	1	1	0	0

Brief about Dataset :

- This is a csv file containing related information of 5172 randomly picked email files and their respective labels for spam or not-spam classification.

Objective of the Project :

- To classify whether an Email is a spam mail or a useful one. A spam mail is a mail that a recipient hasn’t agreed to receive for the reasons that it is useless to the recipient and mostly contains promotional ads or rogue phishing attempts.
- Searching for an important mail in an inbox overflowing with spam mails could be related with searching for a needle in a haystack.
- This ML classification project would enable a user to classify between the two, a spam mail and a useful one and separate them out.

Classifiers Used For Comparison

- Decision Tree Classifier
  - Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems
- Random Forest Classifier
  - Random forests are an ensemble learning method for classification, regression and other tasks.
- K Neighbors Classifier
  - This a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems.

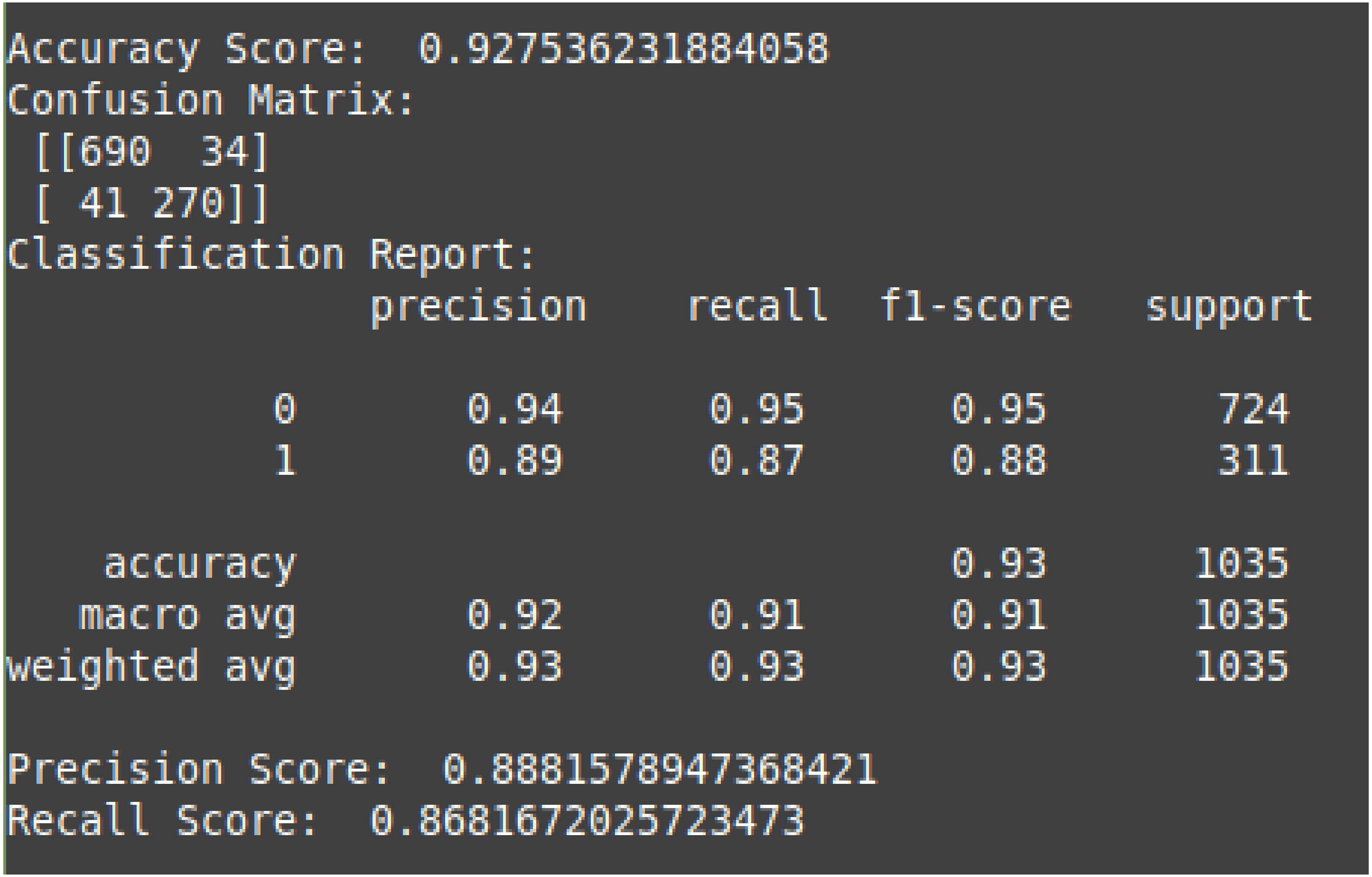
Classifiers Used For Comparison

- Naive Bayes Classifier
  - Naive Bayes classifier is a probabilistic machine learning model that's used for classification task.
- Logistic Regression
  - Logistic regression is basically a supervised classification algorithm.

PREPROCESSING

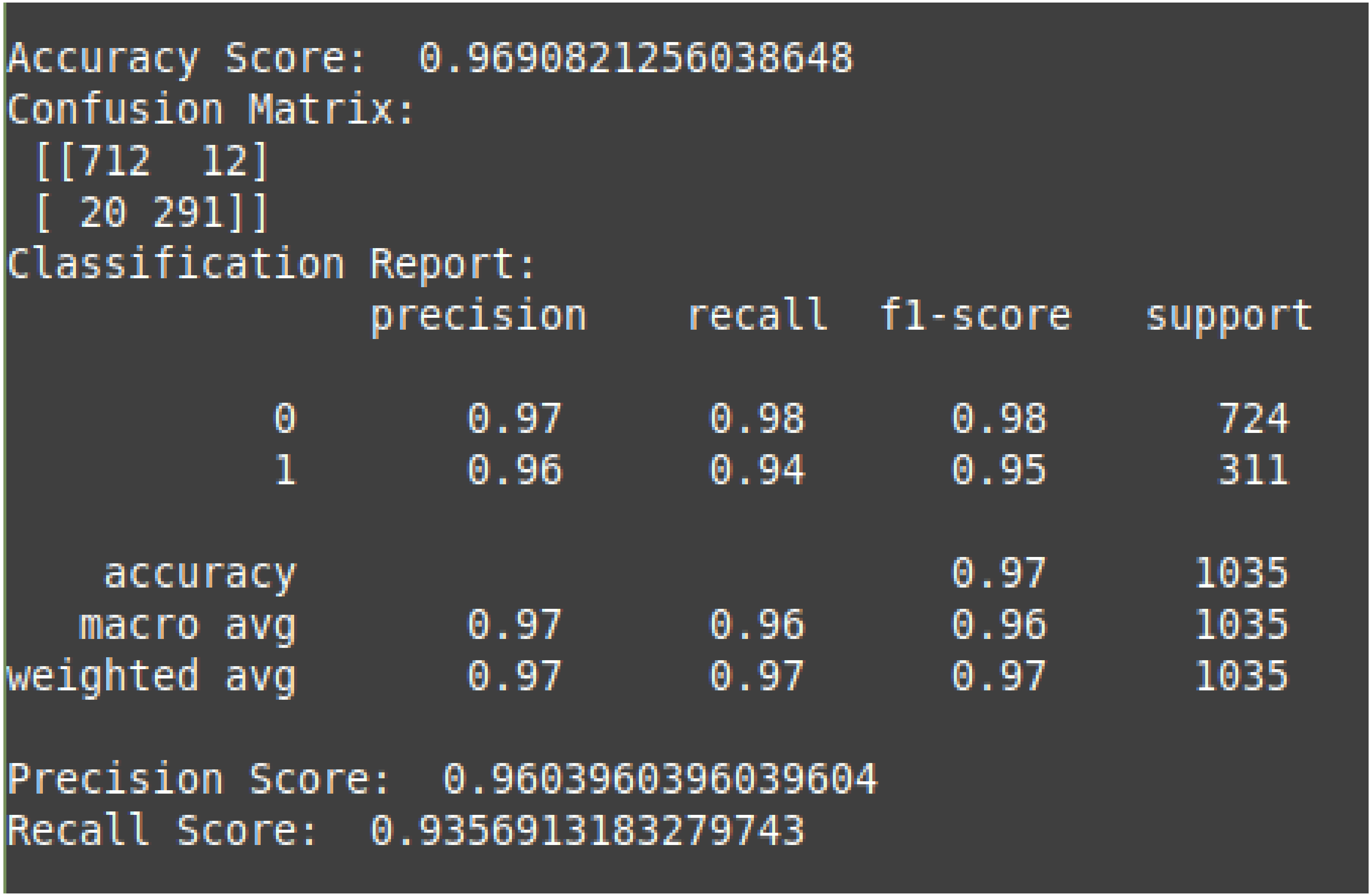
- **Normalization**
  - *Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.*

Decision Tree Classifier



```
Accuracy Score:  0.927536231884058
Confusion Matrix:
 [[690  34]
 [ 41 270]]
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.95      0.95       724
           1       0.89      0.87      0.88       311

    accuracy                           0.93      1035
   macro avg       0.92      0.91      0.91      1035
weighted avg       0.93      0.93      0.93      1035

Precision Score:  0.8881578947368421
Recall Score:  0.8681672025723473
```

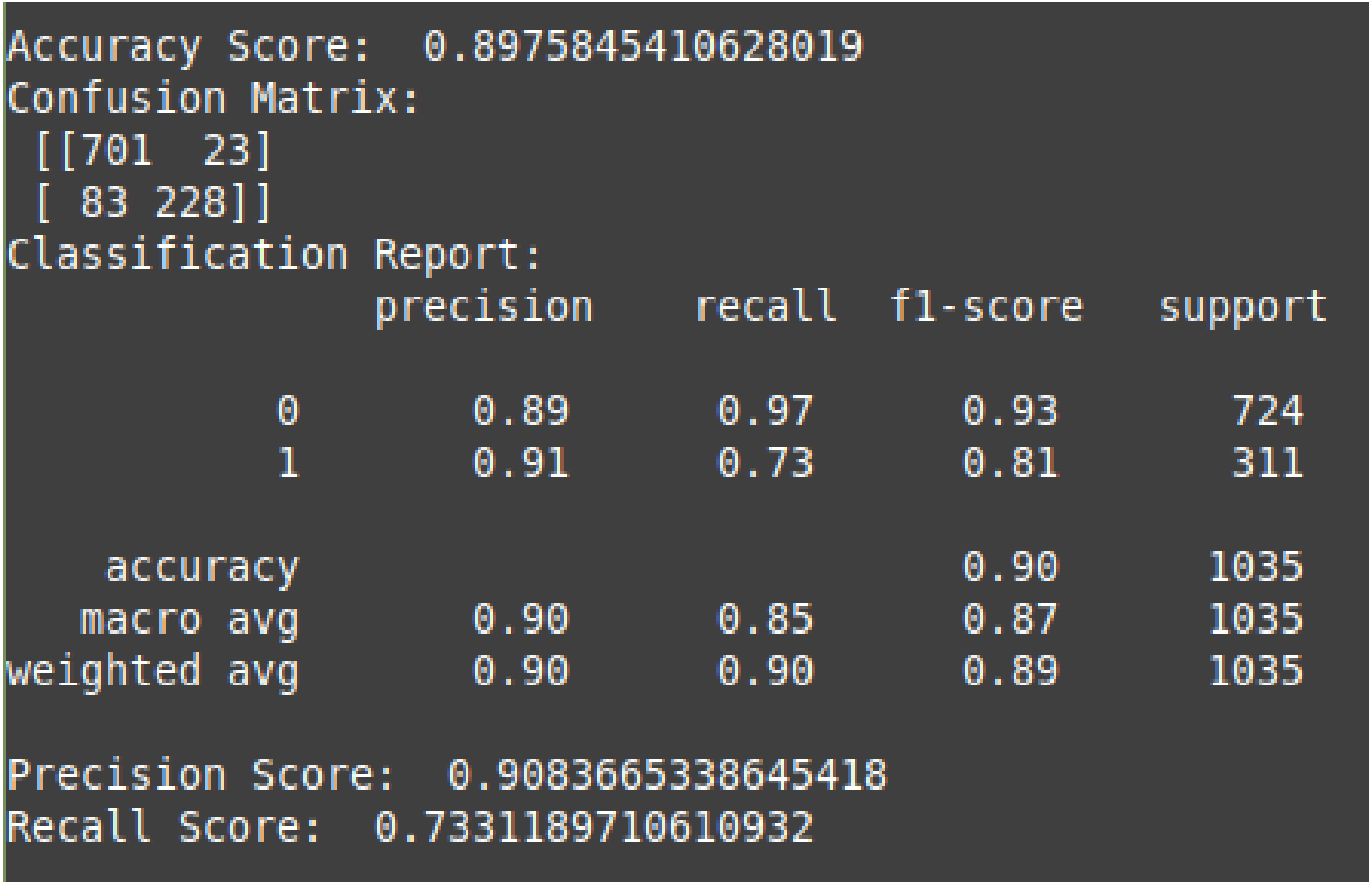
Random Forest Classifier



```
Accuracy Score:  0.9690821256038648
Confusion Matrix:
 [[712  12]
 [ 20 291]]
Classification Report:
               precision    recall  f1-score   support

           0       0.97      0.98      0.98       724
           1       0.96      0.94      0.95       311

    accuracy                           0.97      1035
   macro avg       0.97      0.96      0.96      1035
weighted avg       0.97      0.97      0.97      1035

Precision Score:  0.9603960396039604
Recall Score:  0.9356913183279743
```

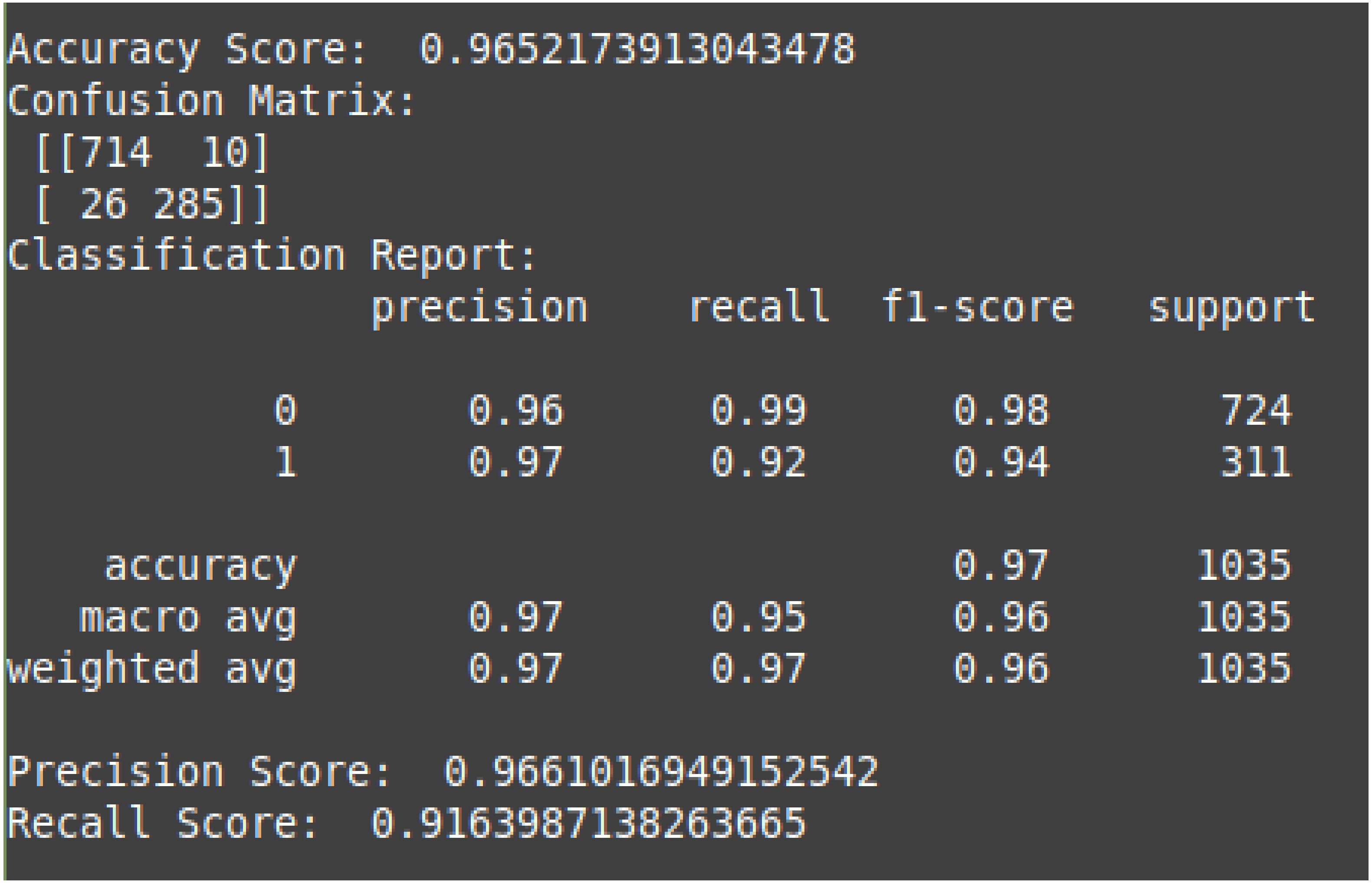
This is the BEST SUITABLE Classifier

K Neighbors Classifier



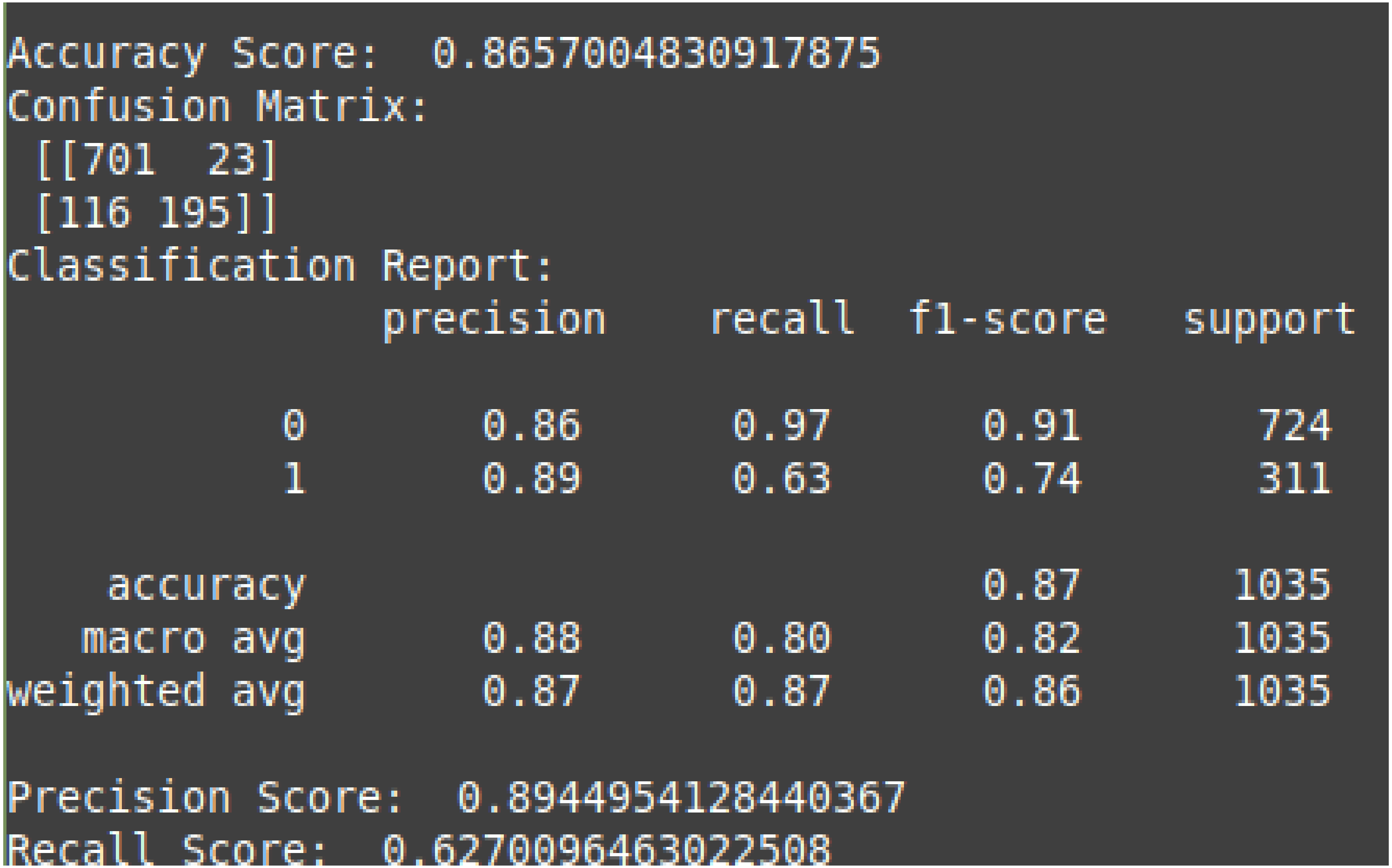
```
Accuracy Score:  0.8975845410628019
Confusion Matrix:
 [[701  23]
 [ 83 228]]
Classification Report:
               precision    recall  f1-score   support

           0       0.89      0.97      0.93       724
           1       0.91      0.73      0.81       311

    accuracy                           0.90      1035
   macro avg       0.90      0.85      0.87      1035
weighted avg       0.90      0.90      0.89      1035

Precision Score:  0.9083665338645418
Recall Score:  0.7331189710610932
```

Naive Bayes Classifier



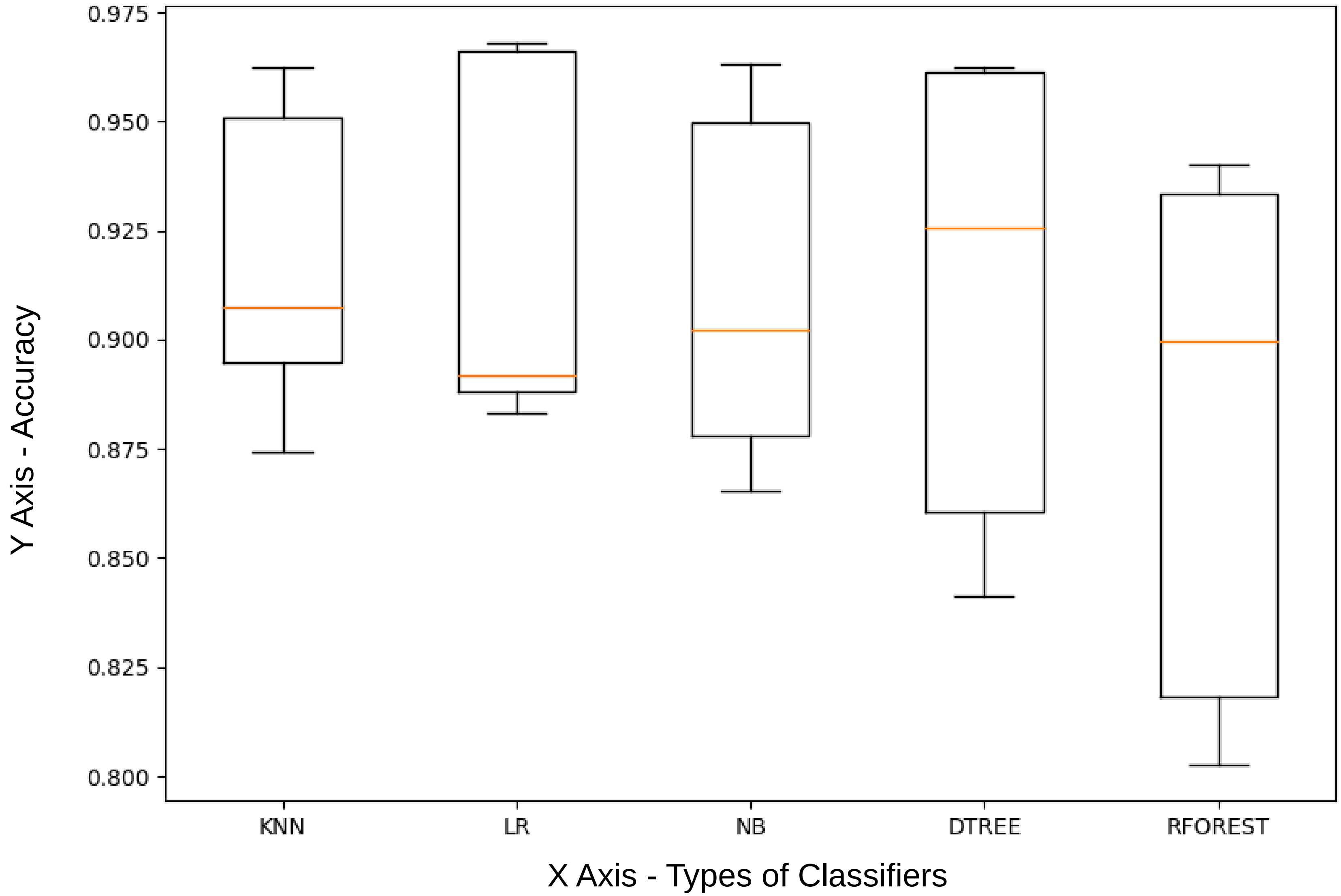
```
Accuracy Score:  0.9652173913043478
Confusion Matrix:
 [[714  10]
 [ 26 285]]
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.98       724
           1       0.97      0.92      0.94       311

    accuracy                           0.97      1035
   macro avg       0.97      0.95      0.96      1035
weighted avg       0.97      0.97      0.96      1035

Precision Score:  0.9661016949152542
Recall Score:  0.9163987138263665
```

Logistic Regression



```
Accuracy Score:  0.8657004830917875
Confusion Matrix:
 [[701  23]
 [116 195]]
Classification Report:
               precision    recall  f1-score   support

           0       0.86      0.97      0.91       724
           1       0.89      0.63      0.74       311

    accuracy                           0.87      1035
   macro avg       0.88      0.80      0.82      1035
weighted avg       0.87      0.87      0.86      1035

Precision Score:  0.8944954128440367
Recall Score:  0.6270096463022508
```

BoxPlot Accuracy Comparison



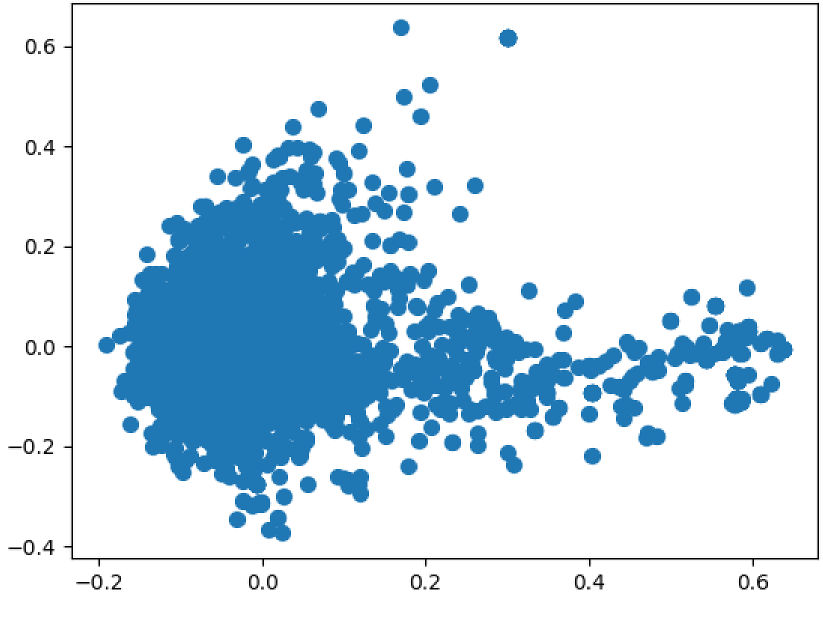
PCA

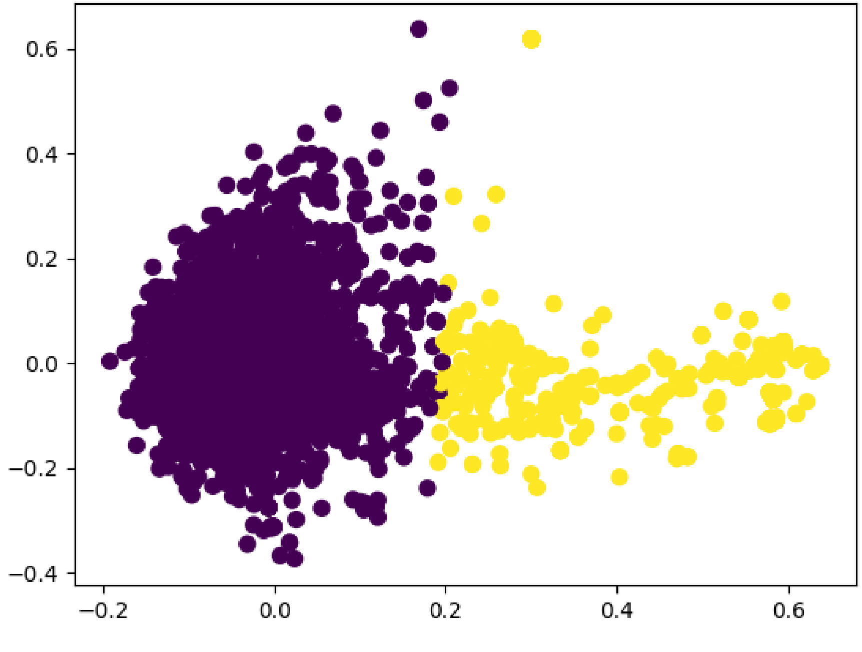
Principal component Analysis dimensionality reduction

- Dimensionality reduction involves reducing the number of input variables or columns in modeling data. PCA is a technique from linear algebra that can be used to automatically perform dimensionality reduction.

Binary Classification depicted using Cluster

- PCA used to reduce 3000 columns to 2 columns.





THANK YOU