DMT4BAS – Homework3

Roesler-Franz, Vukovic

Part1

KNeighborsClassifier

Q1

	Training	Test
Spam	122	53
Ham	120	52

Q2

TfidfVectorizer parameters and values:

Tokenizer: [None, stemming_tokenizer]

Ngram range: (1, 1), (1, 2)

KNeighborsClassifier parameters and values:

N_neighbors: [3, 5, 10]

Leaf_size: [30, 50]

Q3

By performing multiple splits of the datasets during cross-validation and taking an average of each run while increasing performance by setting n_jobs parameter to the value of particular machine number of cores.

Q4

TfidfVectorizer parameters best values:

Tokenizer: None

Ngram range: (1, 1)

KNeighborsClassifier parameters best values:

N_neighbors: 5

Leaf_size: 30

Q5

	Precision	Recall	F1-score	Support
Positive	0.84	0.98	0.90	52
Negative	0.98	0.81	0.89	53
Avg / Total	0.91	0.90	0.89	105

Q6

	Predicted: No	Predicted: Yes
Actual: No	51	1
Actual: Yes	10	43

Q7. Normalized Accuracy Score: **0.89523809523809528**

Q8. The Matthews-Correlation-Coefficient value: **0.80264383325900168**

Part2

KNeighborsClassifier

Q1

	Training	Test
Positive	308	308
Negative	249	250

Q2

TfidfVectorizer parameters and values:

Tokenizer: [None, stemming_tokenizer]

Ngram range: (1, 1), (1, 2)

KNeighborsClassifier parameters and values:

N_neighbors: [3, 5, 10]

Leaf_size: [30, 50]

Q3

By performing multiple splits of the datasets during cross-validation and taking an average of each run while increasing performance by setting n_jobs parameter to the value of particular machine number of cores.

Q4

TfidfVectorizer parameters best values:

Tokenizer: stemming_tokenizer

Ngram range: (1, 2)

KNeighborsClassifier parameters best values:

N_neighbors: 3

Leaf_size: 30

Q5

	Precision	Recall	F1-score	Support
Positive	0.81	0.95	0.88	308
Negative	0.92	0.73	0.81	250
Avg / Total	0.86	0.85	0.85	558

Q6

	Predicted: Positive	Predicted: Negative
Actual: Positive	293	15
Actual: Negative	68	182

Q7. Normalized Accuracy Score: **0.85125448028673834**

Q8. The Matthews-Correlation-Coefficient value: **0.70683725698203992**

linear_model.SGDClassifier

Q1

	Training	Test
Positive	308	308
Negative	249	250

Q2

TfidfVectorizer parameters and values:

Tokenizer: [None, stemming_tokenizer]

Ngram range: (1, 1), (1, 2)

linear_model.SGDClassifier parameters and values:

Loss: ['hinge', 'log', 'squared_loss'],

Penalty: ['l2', 'l1', 'elasticnet']

Q3

By performing multiple splits of the datasets during cross-validation and taking an average of each run while increasing performance by setting n_jobs parameter to the value of particular machine number of cores.

Q4

TfidfVectorizer parameters best values:

Tokenizer: stemming_tokenizer

Ngram range: (1, 2)

linear_model.SGDClassifier parameters best values:

Loss: log

Penalty: l1

Q5

	Precision	Recall	F1-score	Support
Positive	0.96	0.98	0.97	308
Negative	0.98	0.95	0.96	250
Avg / Total	0.97	0.97	0.97	558

Q6

	Predicted: Positive	Predicted: Negative
Actual: Positive	302	6
Actual: Negative	12	238

Q7. Normalized Accuracy Score: **0.967741935483871**

Q8. The Matthews-Correlation-Coefficient value: **0.93485345721554125**

SVC Classifier

Q1

	Training	Test
Positive	308	308
Negative	249	250

Q2

TfidfVectorizer parameters and values:

Tokenizer: [None, stemming_tokenizer]

Ngram range: (1, 1), (1, 2)

SVC Classifier parameters and values:

C: [0.5, 1, 2],

Gamma: [1, 2]

C__kernel: ['rbf', 'linear']

Q3

By performing multiple splits of the datasets during cross-validation and taking an average of each run while increasing performance by setting n_jobs parameter to the value of particular machine number of cores.

Q4

TfidfVectorizer parameters best values:

Tokenizer: stemming_tokenizer

Ngram range: (1, 1)

SVC Classifier parameters best values:

C: 2

Gamma: 1

C__kernel: rbf

Q5

	Precision	Recall	F1-score	Support
Positive	0.94	0.98	0.96	308
Negative	0.97	0.93	0.95	250
Avg / Total	0.96	0.96	0.96	558

Q6

	Predicted: Positive	Predicted: Negative
Actual: Positive	302	6
Actual: Negative	18	232

Q7. Normalized Accuracy Score: **0.956989247311828**

Q8. The Matthews-Correlation-Coefficient value: **0.91351595896852789**