Test/Validation Analysis: Decision Tree: Dataset 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Min\_samples\_split | min\_sample\_leaf | ccp\_alpha | Accuracy |
| Model 1 | 2 | 3 | 1 | 0.5 |
| Model 2 | 4 | 4 | 0 | 0.71 |
| Model 3 | 2 | 1 | 0 | 0.72 |
| Model 4 | 2 | 6 | 0 | 0.716 |
| Model 5 | 4 | 2 | 0 | 0.717 |

Model 3: Precision == 71.57 | Recall == 71.68 | Accuracy == 71.6 | F1 score == 71.62

Test Accuracy : 86.66%

Test/Validation Analysis: Decision Tree: Dataset 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Min\_samples\_split | min\_sample\_leaf | ccp\_alpha | Accuracy |
| Model 1 | 2 | 1 | 0 | 0.711 |
| Model 2 | 2 | 3 | 1 | 0.5 |
| Model 3 | 4 | 2 | 0 | 0.716 |
| Model 4 | 2 | 6 | 0 | 0.719 |
| Model 5 | 4 | 4 | 0 | 0.706 |

Model : Precision == 72.4 | Recall == 71.12 | Accuracy == 71.90 | F1 score == 71.94

Test Accuracy : 76.66%

Test/Validation Analysis : KNN : Dataset 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | n\_neighbors | leaf\_size | p | Accuracy |
| Model 1 | 10 | 36 | 2 | 0.79 |
| Model 2 | 2 | 28 | 2 | 0.753 |
| Model 3 | 5 | 30 | 2 | 0.777 |
| Model 4 | 6 | 32 | 2 | 0.788 |
| Model 5 | 20 | 40 | 2 | 0.799 |

Model 5: Precision == 78.23 | Recall == 82.96 | Accuracy == 79.94 | F1 score == 80.53

Test Accuracy : 76.66%

Test/Validation Analysis : KNN : Dataset2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | n\_neighbors | Leaf\_size | p | Accuracy |
| Model 1 | 20 | 40 | 2 | 0.783 |
| Model 2 | 10 | 36 | 2 | 0.77 |
| Model 3 | 6 | 32 | 2 | 0.766 |
| Model 4 | 60 | 32 | 2 | 0.799 |
| Model 5 | 120 | 80 | 2 | 0.81 |

Model : Precision == 80.7 | Recall == 81.51 | Accuracy == 81.03 | F1 score == 81.1

Test Accuracy : 73.33%

Test/Validation Analysis : Logestic Regression : Dataset 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | C | Intercept\_scalling | Max\_iter | Accuracy |
| Model 1 | 1 | 1 | 10 | 0.873 |
| Model 2 | 1 | 2 | 20 | 0.891 |
| Model 3 | 1 | 1 | 40 | 0.892 |
| Model 4 | 2 | 1 | 30 | 0.893 |
| Model 5 | 1 | 2 | 100 | 0.893 |

Model : Precision == 87.94 | Recall == 91.0 | Accuracy == 89.26 | F1 score == 89.44

Test Accuracy : 86.66%

Test/Validation Analysis : Logestic Regression : Dataset 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | c | Intercept\_scalling | Max\_iter | Accuracy |
| Model 1 | 1 | 1 | 4 | 0.887 |
| Model 2 | 1 | 1 | 10 | 0.882 |
| Model 3 | 1 | 2 | 20 | 0.887 |
| Model 4 | 2 | 2 | 200 | 0.891 |
| Model 5 | 2 | 2 | 10 | 0.873 |

Model : Precision == 88.57 | Recall == 89.56 | Accuracy == 89.01 | F1 score == 89.06

Test Accuracy: 83.33%

Test/Validation Analysis : Gaussain NB : Dataset 1

|  |  |  |
| --- | --- | --- |
|  | Var\_smoothing | Accuracy |
| Model 1 | 1e-9 | 0.756 |
| Model 2 | 1e-8 | 0.757 |
| Model 3 | 100 | 0.771 |
| Model 4 | 1 | 0.807 |
| Model 5 | 1e-2 | 0.815 |

Model : Precision == 81.94 | Recall == 80.92 | Accuracy == 81.54 | F1 score == 81.42

Test Accuracy: 70%

Test/Validation Analysis : Gaussain NB : Dataset 2

|  |  |  |
| --- | --- | --- |
|  | Var\_smoothing | Accuracy |
| Model 1 | 1e-2 | 0.799 |
| Model 2 | 1 | 0.796 |
| Model 3 | 100 | 0.712 |
| Model 4 | 1e-8 | 0.745 |
| Model 5 | 1e-9 | 0.742 |

Model 5: Precision == 77.09 | Recall == 68.89 | Accuracy == 74.24 | F1 score == 72.76

Test Accuracy: 86.66%

Test/Validation Analysis : Multinomial NB : Dataset 1

|  |  |  |  |
| --- | --- | --- | --- |
|  | alpha | Fit\_prior | Accuracy |
| Model 1 | 120 | True | 0.824 |
| Model 2 | 12 | False | 0.856 |
| Model 3 | 12 | True | 0.856 |
| Model 4 | 10 | False | 0.859 |
| Model 5 | 1 | True | 0.864 |

Model 5 : Precision == 86.4 | Recall == 86.4 | Accuracy == 86.4 | F1 score == 86.4

Test accuracy: 80%

Test/Validation Analysis : Multinomial NB : Dataset 2

|  |  |  |  |
| --- | --- | --- | --- |
|  | alpha | Fit\_prior | Accuracy |
| Model 1 | 1 | True | 0.854 |
| Model 2 | 12 | False | 0.847 |
| Model 3 | 10 | False | 0.847 |
| Model 4 | 10 | True | 0.846 |
| Model 5 | 120 | True | 0.829 |

Model 1: Precision == 85.63 | Recall == 85.01 | Accuracy == 85.39 | F1 score == 85.32

Test Accuracy: 80.0%

Test/Validation Analysis : Random Forest : Dataset 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | N\_estimatros | min\_sample\_split | Min\_samples\_leaf | Accuracy |
| Model 1 | 40 | 2 | 1 | 0.841 |
| Model 2 | 100 | 2 | 1 | 0.846 |
| Model 3 | 200 | 2 | 2 | 0.854 |
| Model 4 | 140 | 2 | 1 | 0.855 |
| Model 5 | 160 | 2 | 1 | 0.855 |

Model: Precision == 85.6 | Recall == 85.36 | Accuracy == 85.5 | F1 score == 85.48

Test Accuracy: 90%

Test/Validation Analysis : Random Forest : Dataset 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | N\_estimatros | min\_sample\_split | Min\_samples\_leaf | Accuracy |
| Model 1 | 40 | 2 | 1 | 0.83 |
| Model 2 | 140 | 2 | 1 | 0.853 |
| Model 3 | 100 | 2 | 1 | 0.854 |
| Model 4 | 160 | 2 | 1 | 0.854 |
| Model 5 | 200 | 2 | 2 | 0.856 |

Model: Precision == 85.82 | Recall == 85.56 | Accuracy == 85.72 | F1 score == 85.69

Test Accuracy: 90%

Test/Validation Analysis : SVC: Dataset 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | C | Intercept\_scalling | Max\_iter | Accuracy |
| Model 1 | 1 | 1 | 1000 | 0.88 |
| Model 2 | 1 | 2 | 4 | 0.866 |
| Model 3 | 1 | 1 | 10 | 0.881 |
| Model 4 | 1 | 1 | 4 | 0.885 |
| Model 5 | 1 | 2 | 5 | 0.886 |

Model 5 : Precision == 89.35 | Recall == 87.56 | Accuracy == 88.56 | F1 score == 88.44

Test Accuracy: 89%

Test/Validation

Analysis : SVC : Dataset 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | C | Intercept\_scalling | Max\_iter | Accuracy |
| Model 1 | 1 | 2 | 5 | 0.817 |
| Model 2 | 1 | 1 | 4 | 0.863 |
| Model 3 | 1 | 1 | 1000 | 0.886 |
| Model 4 | 1 | 2 | 4 | 0.886 |
| Model 5 | 1 | 1 | 10 | 0.887 |

Model : Precision == 88.6 | Recall == 88.66 | Accuracy == 88.64 | F1 score == 88.63

Test Accuracy: 86.66%

Dataset labeled datasets

|  |  |  |
| --- | --- | --- |
| Data sets | Text | Sentiment |
| Mixed dataset | 25,000 | Pos 12500 / Neg 12500 |
| Movie review | 40,000 | Pos 19981 / Neg 20019 |