Mail Spam

Jmal Ahmed
Arous Achraf

Plan

Introduction

An “algorithm” in machine learning is a procedure that is run on data to create a machine learning “model.” It is commonly said to be fit on a dataset which means it is applied on the dataset.
There are many types of algorithms with many different functions and purposes.
The three main ones are:

- Regression: Used to make predictions where the output is a continuous value, such as logistic regression.
- Classification: are those algorithms that are used to classify between the categorical values.
- Clustering: Used to group similar items or clustered data points, such as K-Means.

...

First of all, we set 2 dictionaries for the old models and the new models:

```
oldmodels={"KNN": KNeighborsClassifier(n_neighbors=7),
        "DecisionTree":DecisionTreeClassifier(random_state=0)
        }
```

✓ 0.8s Python

```
newmodels={"Logistic Regression": LogisticRegression(random_state=8, solver='lbfgs', max_iter=2000),
        "Linear SVM":LinearSVC(random_state=8, max_iter=3000),
        "RBF SVM":SVC(kernel="rbf", random_state=8),
        "Multi-layer Perceptron Classification": MLPClassifier(hidden_layer_sizes=[20, 20], \
                                                learning_rate='adaptive', random_state=8),
        "MultinominalNB":MultinomialNB(),
        "RandomForestClassifier":RandomForestClassifier()
        }
```

✓ 0.1s Python

...

We use this Function to know the difference between each model and its caracteristics

```
def report(model):
    y_pred=model.fit(X_train, y_train).predict(X_test)
    print(f"Accuracy for {model_name} model : {accuracy_score(y_test, y_pred)}")
    print(classification_report(y_pred,y_test))
    plot_confusion_matrix(model,X_test,y_test)
    plot_precision_recall_curve(model,X_test,y_test)
    plot_roc_curve(model,X_test,y_test)
```

Decision Tree

A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. The model is a form of supervised learning, meaning that the model is trained and tested on a set of data that contains the desired categorization.

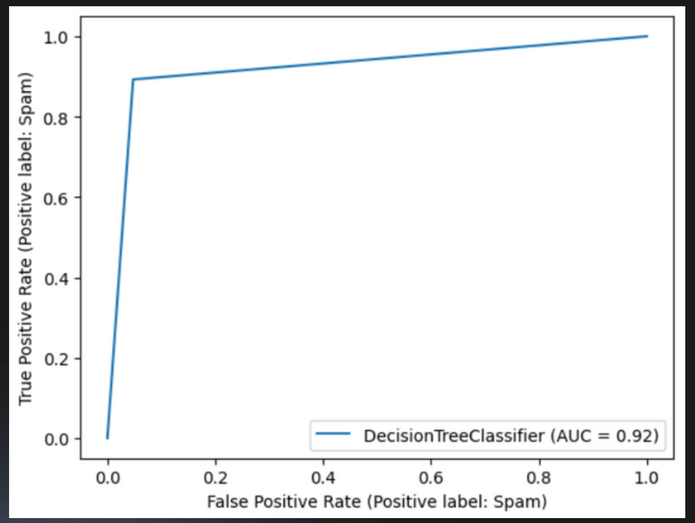
KNN

The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

DecisionTree

	precision	recall	f1-score	support
Not spam	0.95	0.95	0.95	901
Spam	0.89	0.89	0.89	392
accuracy			0.93	1293
macro avg	0.92	0.92	0.92	1293
weighted avg	0.93	0.93	0.93	1293

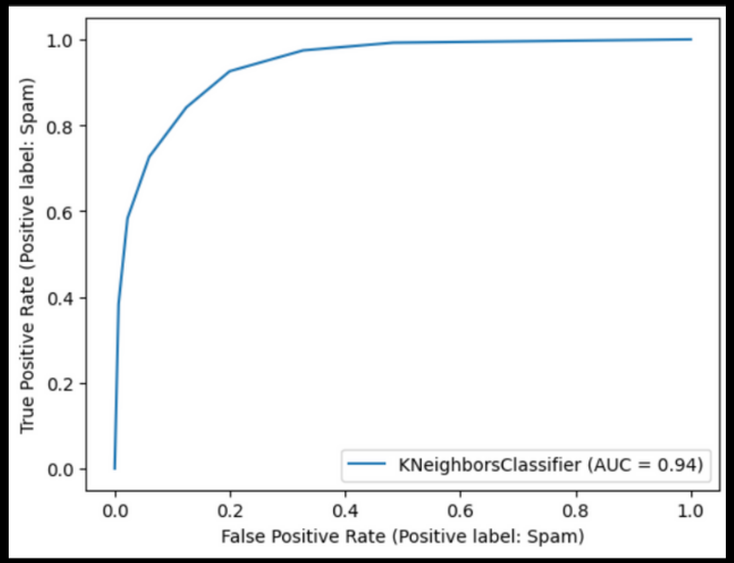
Accuracy for Decision Tree model : 0.93426140 7579273



KNN

	precision	recall	f1-score	support
Not spam	0.95	0.95	0.95	901
Spam	0.89	0.89	0.89	392
accuracy			0.93	1293
macro avg	0.92	0.92	0.92	1293
weighted avg	0.93	0.93	0.93	1293

Accuracy for KNN model : 0.86542923 43387471



LogisticRegression

- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

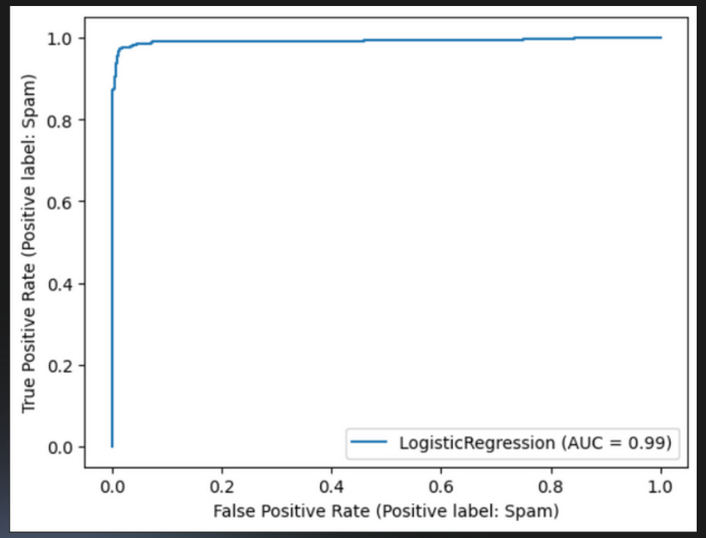
Linear SVM

Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

Logistic Regression

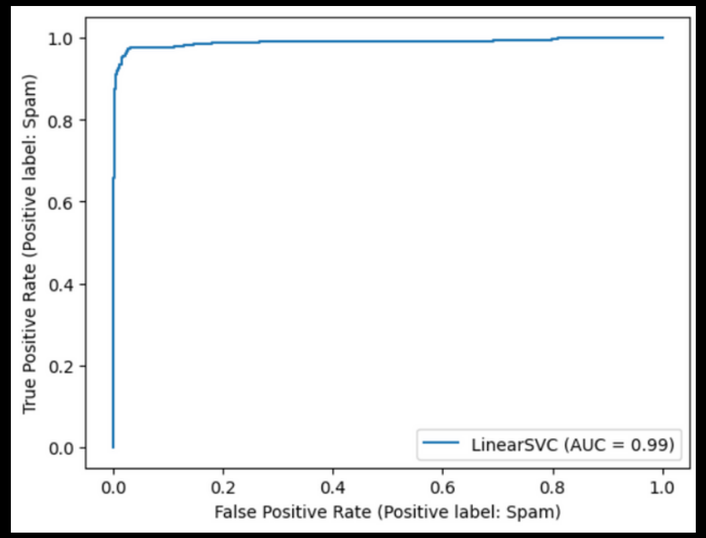
	precision	recall	f1-score	support
Not spam	0.98	0.97	0.98	908
Spam	0.94	0.96	0.95	385
accuracy			0.97	1293
macro avg	0.96	0.97	0.96	1293
weighted avg	0.97	0.97	0.97	1293

Accuracy for Logistic Regression model : 0.98066511 98762568



	precision	recall	f1-score	support
Not spam	0.98	0.97	0.98	908
Spam	0.94	0.96	0.95	385
accuracy			0.97	1293
macro avg	0.96	0.97	0.96	1293
weighted avg	0.97	0.97	0.97	1293

Accuracy for Linear SVM model : 0.97061098 22119103



RBF SVM

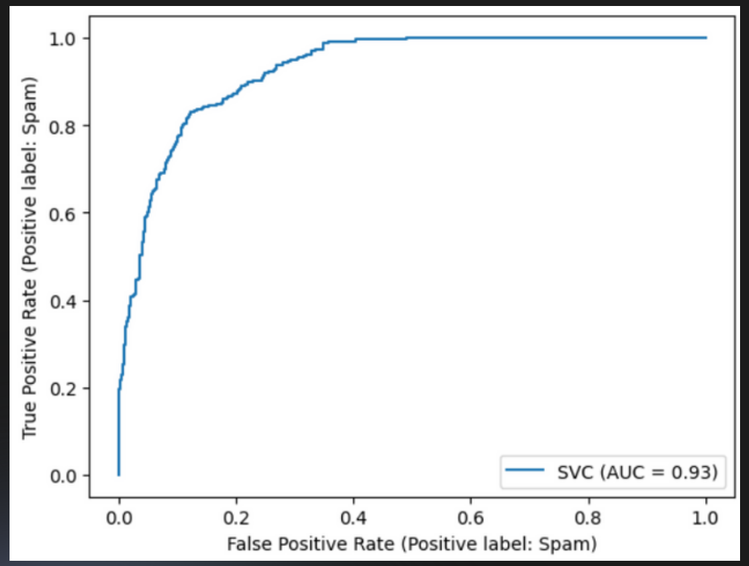
RBF kernels are the most generalized form of kernelization and is one of the most widely used kernels due to its similarity to the Gaussian distribution. The RBF kernel function for two points X_1 and X_2 computes the similarity or how close they are to each other.

Multi-layer Perceptron Classification

A multilayer perceptron (MLP) is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses backpropogation for training the network.

	precision	recall	f1-score	support
Not spam	0.98	0.97	0.98	908
Spam	0.94	0.96	0.95	385
accuracy			0.97	1293
macro avg	0.96	0.97	0.96	1293
weighted avg	0.97	0.97	0.97	1293

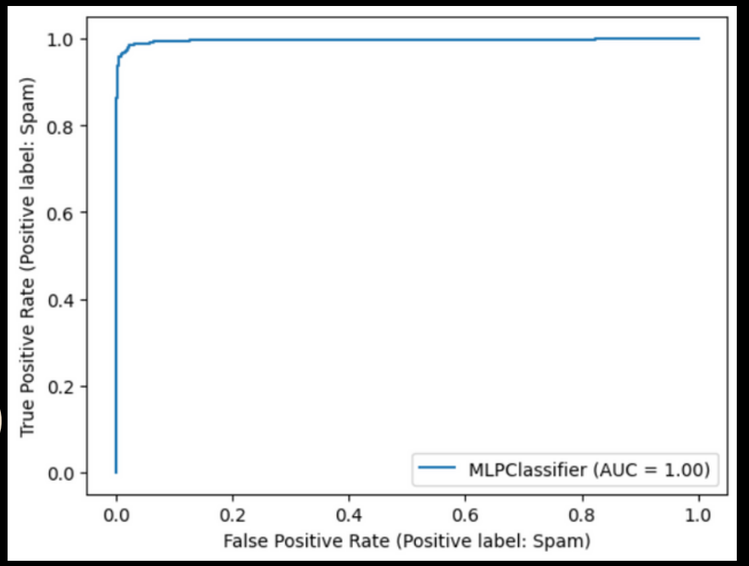
Accuracy for RBF SVM model : 0.7950502706883217



Multi-layer Perceptron Classification

	precision	recall	f1-score	support
Not spam	0.98	0.97	0.98	908
Spam	0.94	0.96	0.95	385
accuracy			0.97	1293
macro avg	0.96	0.97	0.96	1293
weighted avg	0.97	0.97	0.97	1293

Accuracy for Multi-layer Perceptron Classification model : 0.9822119102861562



MultinomialNB

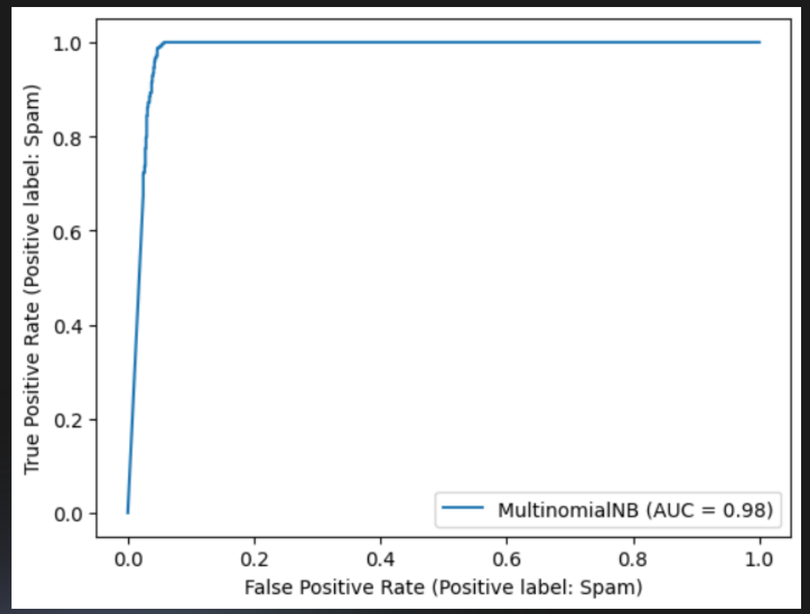
The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance

RandomForestClassifier

Random forest is a supervised learning technique for classification and regression algorithms in machine learning. It's a classifier that combines a number of decision trees on different subsets of a dataset and averages the results to increase the dataset's predicted accuracy.

	precision	recall	f1-score	support
Not spam	0.98	0.97	0.98	908
Spam	0.94	0.96	0.95	385
accuracy			0.97	1293
macro avg	0.96	0.97	0.96	1293
weighted avg	0.97	0.97	0.97	1293

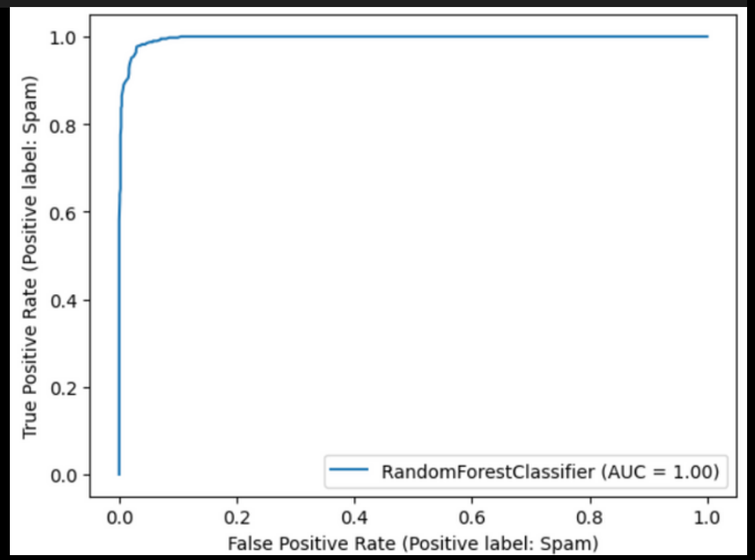
Accuracy for MultinominalNB model : 0.954369682907966



RandomForestClassifier

	precision	recall	f1-score	support
Not spam	0.98	0.97	0.98	908
Spam	0.94	0.96	0.95	385
accuracy			0.97	1293
macro avg	0.96	0.97	0.96	1293
weighted avg	0.97	0.97	0.97	1293

Accuracy for Random Forest Classifier model : 0.9690641918020109



...

For each model, we determined the right parameters (CV=4)

```python
param_grid={ 'max_depth':np.arange(1,101),'criterion':["entropy","gini"]}
grid=GridSearchCV(DecisionTreeClassifier(),param_grid,cv=4)
grid.fit(X,y)
print(grid.best_score_)
print(grid.best_params_)
```

✓ 11m 15.1s Python

```
0.9153132250580047
{'criterion': 'entropy', 'max_depth': 68}
```

```python
param_grid1={ 'max_depth':np.arange(1,101),'criterion':["entropy","gini"]}
grid1=GridSearchCV(RandomForestClassifier(),param_grid1,cv=4)
grid1.fit(X,y)
print(grid1.best_score_)
print(grid1.best_params_)
```

✓ 27m 14.8s Python

```
0.9646171693735499
{'criterion': 'entropy', 'max_depth': 52}
```

...

```
param_grid2={ 'learning_rate':['constant', 'invscaling', 'adaptive'],'activation':['identity', 'logistic']}
grid2=GridSearchCV(MLPClassifier(),param_grid2,cv=4)
grid2.fit(X,y)
print(grid2.best_score_)
print(grid2.best_params_)
```

0.9706109822119103

activation : 'identity'
learning_rate : 'adaptive'

Conclusion

Machine learning is a powerful tool for making predictions from data. However, it is important to remember that machine learning is only as good as the data that is used to train the algorithms. In order to make accurate predictions, it is important to use high-quality data that is representative of the real-world data that the algorithm will be used on.

Thank You
For your
Attention