Machine Learning Model Implementation using Scikit-learn

# 1. Load Dataset

Dataset: SMS Spam Collection.  
The dataset is read from a CSV file using pandas and columns are renamed to 'label' and 'message'.

Code:

df = pd.read\_csv("spam.csv", encoding='latin-1')[['v1', 'v2']]  
df.columns = ['label', 'message']  
df.head()

# 2. Preprocess Dataset

Convert text data to numerical features using CountVectorizer.

Code:

cv = CountVectorizer()  
x = cv.fit\_transform(df['message'])  
y = df['label']

# 3. Split Data into Train and Test

Code:

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size=0.2, random\_state=42)

# 4. Train the Model

Code:

model = MultinomialNB()  
model.fit(x\_train, y\_train)

# 5. Evaluate the Model

Code:

y\_pred = model.predict(x\_test)  
print("Accuracy:", accuracy\_score(y\_test, y\_pred))  
print("Confusion Matrix:", confusion\_matrix(y\_test, y\_pred))  
print("Classification Report:", classification\_report(y\_test, y\_pred))

Results:

Accuracy: 0.978475336322287  
Confusion Matrix:  
[[952 13]  
 [ 11 139]]  
Classification Report:  
 precision recall f1-score support  
 0 0.99 0.99 0.99 965  
 1 0.91 0.93 0.92 150  
 accuracy 0.98 1115  
 macro avg 0.95 0.96 0.95 1115  
weighted avg 0.98 0.98 0.98 1115