

```


from pyspark.sql import SparkSession
from pyspark.ml.feature import StringIndexer, VectorAssembler
from pyspark.ml.classification import NaiveBayes
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
from google.colab import files

```

```

uploaded = files.upload()
for filename in uploaded.keys():
    csv_file = filename

```

 Crime_Data 2.csv

- **Crime_Data 2.csv**(text/csv) - 142968 bytes, last modified: 24/03/2025 - 100% done

Saving Crime_Data 2.csv to Crime_Data 2.csv

```
spark = SparkSession.builder.appName("Crime_NaiveBayes").getOrCreate()
```

```
crime_df = spark.read.csv("Crime_Data 2.csv", header=True, inferSchema=True)
```

```
crime_df = crime_df.select("area", "crime_code", "status").dropna()
```

```

indexer = StringIndexer(inputCol="status", outputCol="label")
crime_df = indexer.fit(crime_df).transform(crime_df)

```

```

assembler = VectorAssembler(inputCols=["area", "crime_code"], outputCol="features")
crime_df = assembler.transform(crime_df)

```

```
train_data, test_data = crime_df.randomSplit([0.8, 0.2], seed=42)
```

```

nb = NaiveBayes(featuresCol="features", labelCol="label", modelType="multinomial")
model = nb.fit(train_data)

```

```
predictions = model.transform(test_data)
```

```

evaluator = MulticlassClassificationEvaluator(labelCol="label", predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)


```

```

last_10 = predictions.select("features", "label", "prediction").tail(10)
for row in last_10:
    print(row)

```

```
print(f"Model Accuracy: {accuracy:.4f}")
```

 Row(features=DenseVector([10.0, 956.0]), label=0.0, prediction=0.0)
 Row(features=DenseVector([11.0, 330.0]), label=0.0, prediction=2.0)
 Row(features=DenseVector([11.0, 745.0]), label=0.0, prediction=0.0)
 Row(features=DenseVector([13.0, 121.0]), label=2.0, prediction=2.0)
 Row(features=DenseVector([15.0, 745.0]), label=0.0, prediction=2.0)
 Row(features=DenseVector([16.0, 420.0]), label=0.0, prediction=2.0)
 Row(features=DenseVector([17.0, 341.0]), label=0.0, prediction=2.0)
 Row(features=DenseVector([17.0, 956.0]), label=0.0, prediction=2.0)
 Row(features=DenseVector([20.0, 310.0]), label=0.0, prediction=2.0)
 Row(features=DenseVector([20.0, 354.0]), label=0.0, prediction=2.0)
 Model Accuracy: 0.8758