task_2_random_forest

February 22, 2026

[1]:
```
# # Mount Google Drive
# from google.colab import drive
# drive.mount('/content/drive')

# Install PySpark
!pip install pyspark

# Imports
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, rand, when
from pyspark.ml.feature import StringIndexer, RFormula, StandardScaler
from pyspark.ml import Pipeline
from pyspark.ml.classification import RandomForestClassifier
from pyspark.ml.tuning import ParamGridBuilder, CrossValidator
from pyspark.ml.evaluation import BinaryClassificationEvaluator,␣
 ↪MulticlassClassificationEvaluator
from pyspark.mllib.evaluation import MulticlassMetrics
from pyspark.sql.functions import col, abs as spark_abs
from pyspark.sql.types import FloatType
import matplotlib.pyplot as plt
import seaborn as sns
import time

# Initialize Spark Session
spark = SparkSession.builder \
    .appName("Task2_RF_AttackLabel") \
    .config("spark.driver.memory", "4g") \
    .getOrCreate()

print("Spark Session Created Successfully!")
```

```
Requirement already satisfied: pyspark in /Users/jju/Documents/SIM/Semester 1,
2026/CSCI316 Big Data Mining/Assignments/.venv/lib/python3.9/site-packages
(3.5.1)
Requirement already satisfied: py4j==0.10.9.7 in
/Users/jju/Documents/SIM/Semester 1, 2026/CSCI316 Big Data
Mining/Assignments/.venv/lib/python3.9/site-packages (from pyspark) (0.10.9.7)
```

```
[notice] A new release of pip is
available: 25.3 -> 26.0.1
[notice] To update, run:
pip install --upgrade pip
```

```
26/02/22 14:18:53 WARN Utils: Your hostname, ijuwon-ui-MacBookPro.local resolves
to a loopback address: 127.0.0.1; using 192.168.0.8 instead (on interface en0)
26/02/22 14:18:53 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
26/02/22 14:18:53 WARN NativeCodeLoader: Unable to load native-hadoop library
for your platform… using builtin-java classes where applicable
26/02/22 14:18:54 WARN Utils: Service 'SparkUI' could not bind on port 4040.
Attempting port 4041.
26/02/22 14:18:54 WARN Utils: Service 'SparkUI' could not bind on port 4041.
Attempting port 4042.
```

```
Spark Session Created Successfully!
```

[2]:
```
def load_and_prep_spark_data():
    train_path = '/content/drive/MyDrive/University/CSCI316 - Big Data Mining␣
↪Techniques/Group Assignment/UNSW_NB15_training-set.csv'
    test_path = '/content/drive/MyDrive/University/CSCI316 - Big Data Mining␣
↪Techniques/Group Assignment/UNSW_NB15_testing-set.csv'

     # Juwon's Local file path
    test_path = '/Users/jju/Documents/SIM/Semester 1, 2026/CSCI316 Big Data␣
↪Mining/Assignments/Group_Assignment_Database/UNSW_NB15_testing-set.csv'
    train_path = '/Users/jju/Documents/SIM/Semester 1, 2026/CSCI316 Big Data␣
↪Mining/Assignments/Group_Assignment_Database/UNSW_NB15_training-set.csv'

    print("Loading data into Spark...")
    df_train_orig = spark.read.csv(train_path, header=True, inferSchema=True)
    df_test_orig = spark.read.csv(test_path, header=True, inferSchema=True)

    # 1. Combine and Drop
    df_full = df_train_orig.unionByName(df_test_orig).drop('id', 'attack_cat')

    df_full = df_full.withColumn("pkt_ratio", (col("spkts") + 1) /␣
↪(col("dpkts") + 1))
    df_full = df_full.withColumn("ttl_gap", spark_abs(col("sttl") -␣
↪col("dttl")))

    # Print Feature Info (Match Task 1 style)
    feature_cols = [c for c in df_full.columns if c != 'label']
    print(f"Features used for training ({len(feature_cols)} total):")
```

```
    print(feature_cols)
    print("-" * 50)

    # 2. Stratified Split
    zeros = df_full.filter(col("label") == 0)
    ones = df_full.filter(col("label") == 1)
    train_0, val_0, test_0 = zeros.randomSplit([0.7, 0.15, 0.15], seed=42)
    train_1, val_1, test_1 = ones.randomSplit([0.7, 0.15, 0.15], seed=42)

    train_data = train_0.union(train_1)
    val_data = val_0.union(val_1)
    test_data = test_0.union(test_1)

    print(f"Data Loaded and Split: Train: {train_data.count()}, Val: {val_data.
↪count()}, Test: {test_data.count()}")
    return train_data, val_data, test_data

train_df, val_df, test_df = load_and_prep_spark_data()
```

```
Loading data into Spark…
Features used for training (44 total):
['dur', 'proto', 'service', 'state', 'spkts', 'dpkts', 'sbytes', 'dbytes',
'rate', 'sttl', 'dttl', 'sload', 'dload', 'sloss', 'dloss', 'sinpkt', 'dinpkt',
'sjit', 'djit', 'swin', 'stcpb', 'dtcpb', 'dwin', 'tcprtt', 'synack', 'ackdat',
'smean', 'dmean', 'trans_depth', 'response_body_len', 'ct_srv_src',
'ct_state_ttl', 'ct_dst_ltm', 'ct_src_dport_ltm', 'ct_dst_sport_ltm',
'ct_dst_src_ltm', 'is_ftp_login', 'ct_ftp_cmd', 'ct_flw_http_mthd',
'ct_src_ltm', 'ct_srv_dst', 'is_sm_ips_ports', 'pkt_ratio', 'ttl_gap']
--------------------------------------------------

26/02/22 14:18:58 WARN SparkStringUtils: Truncated the string representation of
a plan since it was too large. This behavior can be adjusted by setting
'spark.sql.debug.maxToStringFields'.

Data Loaded and Split: Train: 180973, Val: 38573, Test: 38127
```

[3]:
```
def train_spark_rf(train_data, val_data):
# 1. Identify Columns
    categorical_cols = ['proto', 'service', 'state']
    # These are the columns we manually added in load_and_prep_spark_data
    extra_features = ["pkt_ratio", "ttl_gap"]

    # 2. StringIndexer Stage (Manual Ordinal Encoding)
    indexers = [StringIndexer(inputCol=c, outputCol=c+"_idx",␣
↪handleInvalid="keep")
                for c in categorical_cols]

    # 3. Define the Formulas
```

```
    # We use the _idx versions of categorical columns
    indexed_col_names = [c+"_idx" for c in categorical_cols]
    numeric_cols = [c for c, t in train_data.dtypes if t != 'string' and c not␣
↪in (['label'] + extra_features)]

    formula_with = "label ~ " + " + ".join(indexed_col_names + numeric_cols +␣
↪extra_features)
    formula_without = "label ~ " + " + ".join(indexed_col_names + numeric_cols)

    # 4. Setup RFormula and Model
    rf_formula = RFormula(featuresCol="features", labelCol="label_target",␣
↪handleInvalid="keep")
    rf = RandomForestClassifier(labelCol="label_target",␣
↪featuresCol="features", seed=42, maxBins=150)

    # 5. Build the Pipeline
    # IMPORTANT: indexers MUST come before rf_formula
    pipeline = Pipeline(stages=indexers + [rf_formula, rf])

    # 6. Build ParamGrid
    paramGrid = (ParamGridBuilder()
                 .addGrid(rf_formula.formula, [formula_with, formula_without])
                 .addGrid(rf.numTrees, [50, 100])
                 .addGrid(rf.maxDepth, [10, 12])
                 .build())

    # 7. Evaluator and CrossValidator
    evaluator = MulticlassClassificationEvaluator(labelCol="label_target",␣
↪metricName="f1")
    cv = CrossValidator(estimator=pipeline,
                        estimatorParamMaps=paramGrid,
                        evaluator=evaluator,
                        numFolds=3)

    print("Starting Grid Search (Comparing Extra Features)...")
    start_time = time.time()
    cvModel = cv.fit(train_data)
    duration = time.time() - start_time

    # 5. Extract Results
    avg_metrics = cvModel.avgMetrics
    params = cv.getEstimatorParamMaps()
    results = sorted(zip(params, avg_metrics), key=lambda x: x[1], reverse=True)

    # Print Task 1 style summary
    print(f"\nGrid Search Complete in {duration:.2f} seconds.")
```

```
    best_p = results[0][0]
    has_extra = "True" if "pkt_ratio" in best_p[rf_formula.formula] else "False"

    print(f"Best Parameters: {{'use_extra_features': {has_extra}, 'numTrees':␣
 ↪{best_p[rf.numTrees]}, 'maxDepth': {best_p[rf.maxDepth]}}}")
    print(f"Validation F1-Score: {max(avg_metrics):.4f}")

    return cvModel.bestModel

best_rf_model = train_spark_rf(train_df, val_df)
```

```
Starting Grid Search (Comparing Extra Features)…
26/02/22 14:19:11 WARN DAGScheduler: Broadcasting large task binary with size
1268.0 KiB
26/02/22 14:19:12 WARN DAGScheduler: Broadcasting large task binary with size
1766.0 KiB
26/02/22 14:19:13 WARN DAGScheduler: Broadcasting large task binary with size
2.4 MiB
26/02/22 14:19:16 WARN DAGScheduler: Broadcasting large task binary with size
1726.6 KiB
26/02/22 14:19:27 WARN DAGScheduler: Broadcasting large task binary with size
1268.0 KiB
26/02/22 14:19:29 WARN DAGScheduler: Broadcasting large task binary with size
1766.0 KiB
26/02/22 14:19:30 WARN DAGScheduler: Broadcasting large task binary with size
2.4 MiB
26/02/22 14:19:32 WARN DAGScheduler: Broadcasting large task binary with size
3.3 MiB
26/02/22 14:19:37 WARN DAGScheduler: Broadcasting large task binary with size
4.6 MiB
26/02/22 14:19:43 WARN DAGScheduler: Broadcasting large task binary with size
3.1 MiB
26/02/22 14:19:53 WARN DAGScheduler: Broadcasting large task binary with size
1024.3 KiB
26/02/22 14:19:56 WARN DAGScheduler: Broadcasting large task binary with size
1462.4 KiB
26/02/22 14:19:59 WARN DAGScheduler: Broadcasting large task binary with size
2.1 MiB
26/02/22 14:20:03 WARN DAGScheduler: Broadcasting large task binary with size
3.0 MiB
26/02/22 14:20:09 WARN DAGScheduler: Broadcasting large task binary with size
4.4 MiB
26/02/22 14:20:19 WARN DAGScheduler: Broadcasting large task binary with size
3.1 MiB
26/02/22 14:20:28 WARN DAGScheduler: Broadcasting large task binary with size
1024.3 KiB
26/02/22 14:20:30 WARN DAGScheduler: Broadcasting large task binary with size
```

```
1462.4 KiB
26/02/22 14:20:33 WARN DAGScheduler: Broadcasting large task binary with size
2.1 MiB
26/02/22 14:20:37 WARN DAGScheduler: Broadcasting large task binary with size
3.0 MiB
26/02/22 14:20:42 WARN DAGScheduler: Broadcasting large task binary with size
4.4 MiB
26/02/22 14:20:51 WARN DAGScheduler: Broadcasting large task binary with size
6.4 MiB
26/02/22 14:21:05 WARN DAGScheduler: Broadcasting large task binary with size
9.0 MiB
26/02/22 14:21:22 WARN DAGScheduler: Broadcasting large task binary with size
1238.8 KiB
26/02/22 14:21:32 WARN DAGScheduler: Broadcasting large task binary with size
5.9 MiB
26/02/22 14:21:45 WARN DAGScheduler: Broadcasting large task binary with size
1226.6 KiB
26/02/22 14:21:48 WARN DAGScheduler: Broadcasting large task binary with size
1707.7 KiB
26/02/22 14:21:51 WARN DAGScheduler: Broadcasting large task binary with size
2.3 MiB
26/02/22 14:21:56 WARN DAGScheduler: Broadcasting large task binary with size
1707.8 KiB
26/02/22 14:22:10 WARN DAGScheduler: Broadcasting large task binary with size
1226.6 KiB
26/02/22 14:22:13 WARN DAGScheduler: Broadcasting large task binary with size
1707.7 KiB
26/02/22 14:22:16 WARN DAGScheduler: Broadcasting large task binary with size
2.3 MiB
26/02/22 14:22:20 WARN DAGScheduler: Broadcasting large task binary with size
3.3 MiB
26/02/22 14:22:26 WARN DAGScheduler: Broadcasting large task binary with size
4.5 MiB
26/02/22 14:22:33 WARN DAGScheduler: Broadcasting large task binary with size
3.1 MiB
26/02/22 14:22:46 WARN DAGScheduler: Broadcasting large task binary with size
1003.9 KiB
26/02/22 14:22:49 WARN DAGScheduler: Broadcasting large task binary with size
1437.9 KiB
26/02/22 14:22:53 WARN DAGScheduler: Broadcasting large task binary with size
2.1 MiB
26/02/22 14:22:58 WARN DAGScheduler: Broadcasting large task binary with size
3.0 MiB
26/02/22 14:23:05 WARN DAGScheduler: Broadcasting large task binary with size
4.3 MiB
26/02/22 14:23:15 WARN DAGScheduler: Broadcasting large task binary with size
3.0 MiB
26/02/22 14:23:27 WARN DAGScheduler: Broadcasting large task binary with size
```

```
1003.9 KiB
26/02/22 14:23:31 WARN DAGScheduler: Broadcasting large task binary with size
1437.9 KiB
26/02/22 14:23:36 WARN DAGScheduler: Broadcasting large task binary with size
2.1 MiB
26/02/22 14:23:41 WARN DAGScheduler: Broadcasting large task binary with size
3.0 MiB
26/02/22 14:23:49 WARN DAGScheduler: Broadcasting large task binary with size
4.3 MiB
26/02/22 14:23:56 WARN DAGScheduler: Broadcasting large task binary with size
6.2 MiB
26/02/22 14:24:09 WARN DAGScheduler: Broadcasting large task binary with size
8.7 MiB
26/02/22 14:24:19 WARN DAGScheduler: Broadcasting large task binary with size
1187.3 KiB
26/02/22 14:24:28 WARN DAGScheduler: Broadcasting large task binary with size
5.8 MiB
26/02/22 14:24:42 WARN DAGScheduler: Broadcasting large task binary with size
1274.8 KiB
26/02/22 14:24:44 WARN DAGScheduler: Broadcasting large task binary with size
1774.6 KiB
26/02/22 14:24:47 WARN DAGScheduler: Broadcasting large task binary with size
2.4 MiB
26/02/22 14:24:52 WARN DAGScheduler: Broadcasting large task binary with size
1749.5 KiB
26/02/22 14:25:02 WARN DAGScheduler: Broadcasting large task binary with size
1274.8 KiB
26/02/22 14:25:03 WARN DAGScheduler: Broadcasting large task binary with size
1774.6 KiB
26/02/22 14:25:05 WARN DAGScheduler: Broadcasting large task binary with size
2.4 MiB
26/02/22 14:25:08 WARN DAGScheduler: Broadcasting large task binary with size
3.4 MiB
26/02/22 14:25:12 WARN DAGScheduler: Broadcasting large task binary with size
4.7 MiB
26/02/22 14:25:20 WARN DAGScheduler: Broadcasting large task binary with size
3.2 MiB
26/02/22 14:25:26 WARN DAGScheduler: Broadcasting large task binary with size
1022.0 KiB
26/02/22 14:25:28 WARN DAGScheduler: Broadcasting large task binary with size
1472.8 KiB
26/02/22 14:25:30 WARN DAGScheduler: Broadcasting large task binary with size
2.1 MiB
26/02/22 14:25:33 WARN DAGScheduler: Broadcasting large task binary with size
3.1 MiB
26/02/22 14:25:38 WARN DAGScheduler: Broadcasting large task binary with size
4.4 MiB
26/02/22 14:25:45 WARN DAGScheduler: Broadcasting large task binary with size
```

```
3.0 MiB
26/02/22 14:25:52 WARN DAGScheduler: Broadcasting large task binary with size
1022.0 KiB
26/02/22 14:25:54 WARN DAGScheduler: Broadcasting large task binary with size
1472.8 KiB
26/02/22 14:25:56 WARN DAGScheduler: Broadcasting large task binary with size
2.1 MiB
26/02/22 14:25:59 WARN DAGScheduler: Broadcasting large task binary with size
3.1 MiB
26/02/22 14:26:04 WARN DAGScheduler: Broadcasting large task binary with size
4.4 MiB
26/02/22 14:26:14 WARN DAGScheduler: Broadcasting large task binary with size
6.3 MiB
26/02/22 14:26:28 WARN DAGScheduler: Broadcasting large task binary with size
8.9 MiB
26/02/22 14:26:41 WARN DAGScheduler: Broadcasting large task binary with size
1199.9 KiB
26/02/22 14:26:48 WARN DAGScheduler: Broadcasting large task binary with size
5.8 MiB
26/02/22 14:26:59 WARN DAGScheduler: Broadcasting large task binary with size
1243.9 KiB
26/02/22 14:27:00 WARN DAGScheduler: Broadcasting large task binary with size
1734.9 KiB
26/02/22 14:27:02 WARN DAGScheduler: Broadcasting large task binary with size
2.4 MiB
26/02/22 14:27:06 WARN DAGScheduler: Broadcasting large task binary with size
1703.3 KiB
26/02/22 14:27:14 WARN DAGScheduler: Broadcasting large task binary with size
1243.9 KiB
26/02/22 14:27:15 WARN DAGScheduler: Broadcasting large task binary with size
1734.9 KiB
26/02/22 14:27:17 WARN DAGScheduler: Broadcasting large task binary with size
2.4 MiB
26/02/22 14:27:19 WARN DAGScheduler: Broadcasting large task binary with size
3.3 MiB
26/02/22 14:27:23 WARN DAGScheduler: Broadcasting large task binary with size
4.6 MiB
26/02/22 14:27:29 WARN DAGScheduler: Broadcasting large task binary with size
3.1 MiB
26/02/22 14:27:35 WARN DAGScheduler: Broadcasting large task binary with size
1004.4 KiB
26/02/22 14:27:36 WARN DAGScheduler: Broadcasting large task binary with size
1442.2 KiB
26/02/22 14:27:38 WARN DAGScheduler: Broadcasting large task binary with size
2.1 MiB
26/02/22 14:27:41 WARN DAGScheduler: Broadcasting large task binary with size
3.0 MiB
26/02/22 14:27:47 WARN DAGScheduler: Broadcasting large task binary with size
```

```
4.3 MiB
26/02/22 14:27:57 WARN DAGScheduler: Broadcasting large task binary with size
3.0 MiB
26/02/22 14:28:10 WARN DAGScheduler: Broadcasting large task binary with size
1004.4 KiB
26/02/22 14:28:11 WARN DAGScheduler: Broadcasting large task binary with size
1442.2 KiB
26/02/22 14:28:14 WARN DAGScheduler: Broadcasting large task binary with size
2.1 MiB
26/02/22 14:28:18 WARN DAGScheduler: Broadcasting large task binary with size
3.0 MiB
26/02/22 14:28:23 WARN DAGScheduler: Broadcasting large task binary with size
4.3 MiB
26/02/22 14:28:31 WARN DAGScheduler: Broadcasting large task binary with size
6.2 MiB
26/02/22 14:28:52 WARN DAGScheduler: Broadcasting large task binary with size
8.7 MiB
26/02/22 14:29:11 WARN DAGScheduler: Broadcasting large task binary with size
1188.3 KiB
26/02/22 14:29:16 WARN DAGScheduler: Broadcasting large task binary with size
5.7 MiB
26/02/22 14:29:27 WARN DAGScheduler: Broadcasting large task binary with size
1279.8 KiB
26/02/22 14:29:28 WARN DAGScheduler: Broadcasting large task binary with size
1785.1 KiB
26/02/22 14:29:32 WARN DAGScheduler: Broadcasting large task binary with size
2.4 MiB
26/02/22 14:29:37 WARN DAGScheduler: Broadcasting large task binary with size
1751.2 KiB
26/02/22 14:29:47 WARN DAGScheduler: Broadcasting large task binary with size
1279.8 KiB
26/02/22 14:29:49 WARN DAGScheduler: Broadcasting large task binary with size
1785.1 KiB
26/02/22 14:29:52 WARN DAGScheduler: Broadcasting large task binary with size
2.4 MiB
26/02/22 14:29:56 WARN DAGScheduler: Broadcasting large task binary with size
3.4 MiB
26/02/22 14:30:00 WARN DAGScheduler: Broadcasting large task binary with size
4.7 MiB
26/02/22 14:30:08 WARN DAGScheduler: Broadcasting large task binary with size
3.2 MiB
26/02/22 14:30:17 WARN DAGScheduler: Broadcasting large task binary with size
1024.4 KiB
26/02/22 14:30:20 WARN DAGScheduler: Broadcasting large task binary with size
1477.5 KiB
26/02/22 14:30:24 WARN DAGScheduler: Broadcasting large task binary with size
2.1 MiB
26/02/22 14:30:29 WARN DAGScheduler: Broadcasting large task binary with size
```

```
3.1 MiB
26/02/22 14:30:41 WARN DAGScheduler: Broadcasting large task binary with size
4.5 MiB
26/02/22 14:30:53 WARN DAGScheduler: Broadcasting large task binary with size
3.1 MiB
26/02/22 14:31:04 WARN DAGScheduler: Broadcasting large task binary with size
1024.4 KiB
26/02/22 14:31:06 WARN DAGScheduler: Broadcasting large task binary with size
1477.5 KiB
26/02/22 14:31:08 WARN DAGScheduler: Broadcasting large task binary with size
2.1 MiB
26/02/22 14:31:12 WARN DAGScheduler: Broadcasting large task binary with size
3.1 MiB
26/02/22 14:31:19 WARN DAGScheduler: Broadcasting large task binary with size
4.5 MiB
26/02/22 14:31:31 WARN DAGScheduler: Broadcasting large task binary with size
6.4 MiB
26/02/22 14:31:49 WARN DAGScheduler: Broadcasting large task binary with size
9.0 MiB
26/02/22 14:32:05 WARN DAGScheduler: Broadcasting large task binary with size
1210.9 KiB
26/02/22 14:32:11 WARN DAGScheduler: Broadcasting large task binary with size
6.0 MiB
26/02/22 14:32:20 WARN DAGScheduler: Broadcasting large task binary with size
1252.4 KiB
26/02/22 14:32:23 WARN DAGScheduler: Broadcasting large task binary with size
1756.3 KiB
26/02/22 14:32:24 WARN DAGScheduler: Broadcasting large task binary with size
2.4 MiB
26/02/22 14:32:29 WARN DAGScheduler: Broadcasting large task binary with size
1744.1 KiB
26/02/22 14:32:38 WARN DAGScheduler: Broadcasting large task binary with size
1252.4 KiB
26/02/22 14:32:40 WARN DAGScheduler: Broadcasting large task binary with size
1756.3 KiB
26/02/22 14:32:43 WARN DAGScheduler: Broadcasting large task binary with size
2.4 MiB
26/02/22 14:32:46 WARN DAGScheduler: Broadcasting large task binary with size
3.4 MiB
26/02/22 14:32:50 WARN DAGScheduler: Broadcasting large task binary with size
4.7 MiB
26/02/22 14:32:56 WARN DAGScheduler: Broadcasting large task binary with size
3.1 MiB
26/02/22 14:33:02 WARN DAGScheduler: Broadcasting large task binary with size
1007.7 KiB
26/02/22 14:33:04 WARN DAGScheduler: Broadcasting large task binary with size
1446.0 KiB
26/02/22 14:33:07 WARN DAGScheduler: Broadcasting large task binary with size
```

```
2.1 MiB
26/02/22 14:33:09 WARN DAGScheduler: Broadcasting large task binary with size
3.0 MiB
26/02/22 14:33:11 WARN DAGScheduler: Broadcasting large task binary with size
4.3 MiB
26/02/22 14:33:15 WARN DAGScheduler: Broadcasting large task binary with size
3.0 MiB
26/02/22 14:33:19 WARN DAGScheduler: Broadcasting large task binary with size
1007.7 KiB
26/02/22 14:33:20 WARN DAGScheduler: Broadcasting large task binary with size
1446.0 KiB
26/02/22 14:33:21 WARN DAGScheduler: Broadcasting large task binary with size
2.1 MiB
26/02/22 14:33:23 WARN DAGScheduler: Broadcasting large task binary with size
3.0 MiB
26/02/22 14:33:26 WARN DAGScheduler: Broadcasting large task binary with size
4.3 MiB
26/02/22 14:33:30 WARN DAGScheduler: Broadcasting large task binary with size
6.1 MiB
26/02/22 14:33:37 WARN DAGScheduler: Broadcasting large task binary with size
8.5 MiB
26/02/22 14:33:42 WARN DAGScheduler: Broadcasting large task binary with size
1152.0 KiB
26/02/22 14:33:44 WARN DAGScheduler: Broadcasting large task binary with size
5.6 MiB
26/02/22 14:33:54 WARN DAGScheduler: Broadcasting large task binary with size
1276.6 KiB
26/02/22 14:33:55 WARN DAGScheduler: Broadcasting large task binary with size
1795.6 KiB
26/02/22 14:33:56 WARN DAGScheduler: Broadcasting large task binary with size
2.5 MiB
26/02/22 14:33:58 WARN DAGScheduler: Broadcasting large task binary with size
3.5 MiB
26/02/22 14:34:02 WARN DAGScheduler: Broadcasting large task binary with size
4.9 MiB


Grid Search Complete in 904.31 seconds.
Best Parameters: {'use_extra_features': False, 'numTrees': 50, 'maxDepth': 12}
Validation F1-Score: 0.9388
```

[4]:
```python
def evaluate_spark_model(model, test_data):
    print("\nFINAL EVALUATION: UNSEEN TEST DATA (SPARK)")

    # --- FIX: Drop columns if they exist to prevent "already exists" error ---
    # This allows the model's internal pipeline to create them fresh
    cols_to_drop = ["proto_idx", "service_idx", "state_idx", "label_target",␣
 ↪"features"]
```

```
    for c in cols_to_drop:
        if c in test_data.columns:
            test_data = test_data.drop(c)
    # ------------------------------------------------------------------------

    # Make predictions
    predictions = model.transform(test_data)

    # Convert to RDD for MulticlassMetrics
    # Note: RFormula inside the model creates 'label_target'
    predictionAndLabels = predictions.select(
        col("prediction").cast(FloatType()),
        col("label_target").cast(FloatType())
    ).rdd

    metrics = MulticlassMetrics(predictionAndLabels)

    # Overall (Weighted) Metrics
    print(f"\n--- OVERALL PERFORMANCE (Weighted) ---")
    print(f"Accuracy:           {metrics.accuracy:.6f}")
    print(f"Weighted Precision: {metrics.weightedPrecision:.6f}")
    print(f"Weighted Recall:    {metrics.weightedRecall:.6f}")
    print(f"Weighted F1-Score:  {metrics.weightedFMeasure():.6f}")

    # Confusion Matrix
    cm = metrics.confusionMatrix().toArray()
    plt.figure(figsize=(7, 5))
    sns.heatmap(cm, annot=True, fmt='.0f', cmap='Blues',
                xticklabels=['Normal', 'Attack'],
                yticklabels=['Normal', 'Attack'])
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.title('Confusion Matrix - Spark RF')
    plt.show()

# EXECUTION CALL
# Since the function is now "safe", you can pass either test_df or test_indexed
evaluate_spark_model(best_rf_model, test_df)
```

```
FINAL EVALUATION: UNSEEN TEST DATA (SPARK)

/Users/jju/Documents/SIM/Semester 1, 2026/CSCI316 Big Data
Mining/Assignments/.venv/lib/python3.9/site-packages/pyspark/sql/context.py:158:
FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate()
instead.
  warnings.warn(
26/02/22 14:34:06 WARN DAGScheduler: Broadcasting large task binary with size
```

```
3.2 MiB
26/02/22 14:34:07 WARN DAGScheduler: Broadcasting large task binary with size
3.2 MiB


--- OVERALL PERFORMANCE (Weighted) ---

[Stage 1020:==============================================>      (21 + 3) / 24]

Accuracy:           0.937944
Weighted Precision: 0.938256
Weighted Recall:    0.937944
Weighted F1-Score:  0.938060
```