from pyspark.sql import SparkSession

spark = SparkSession.builder.getOrCreate()

df = spark.read.csv(“D:/OneDrive/Desktop/LPU/cluster computing/MLlib/iris.csv”,header=True,inferSchema=True)

df.show()

from pyspark.ml.feature import VectorAssembler

vector\_assembler = VectorAssembler(inputCols=df.columns[:-1],outputCol="features")

df1 = vector\_assembler.transform(df)

from pyspark.ml.feature import StringIndexer

l\_indexer = StringIndexer(inputCol="species", outputCol="labelIndex")

df1 = l\_indexer.fit(df1).transform(df1)

(trainingData, testData) = df1.randomSplit([0.7, 0.3])

from pyspark.ml.classification import DecisionTreeClassifier

from pyspark.ml.evaluation import MulticlassClassificationEvaluator

dt = DecisionTreeClassifier(labelCol="labelIndex", featuresCol="features")

model = dt.fit(trainingData)

predictions = model.transform(testData)

predictions.select("prediction", "labelIndex").show(5)

evaluator = MulticlassClassificationEvaluator(labelCol="labelIndex", predictionCol="prediction",metricName="accuracy")

accuracy = evaluator.evaluate(predictions)

print("Test Error = %g " % (1.0 - accuracy))

print(model)