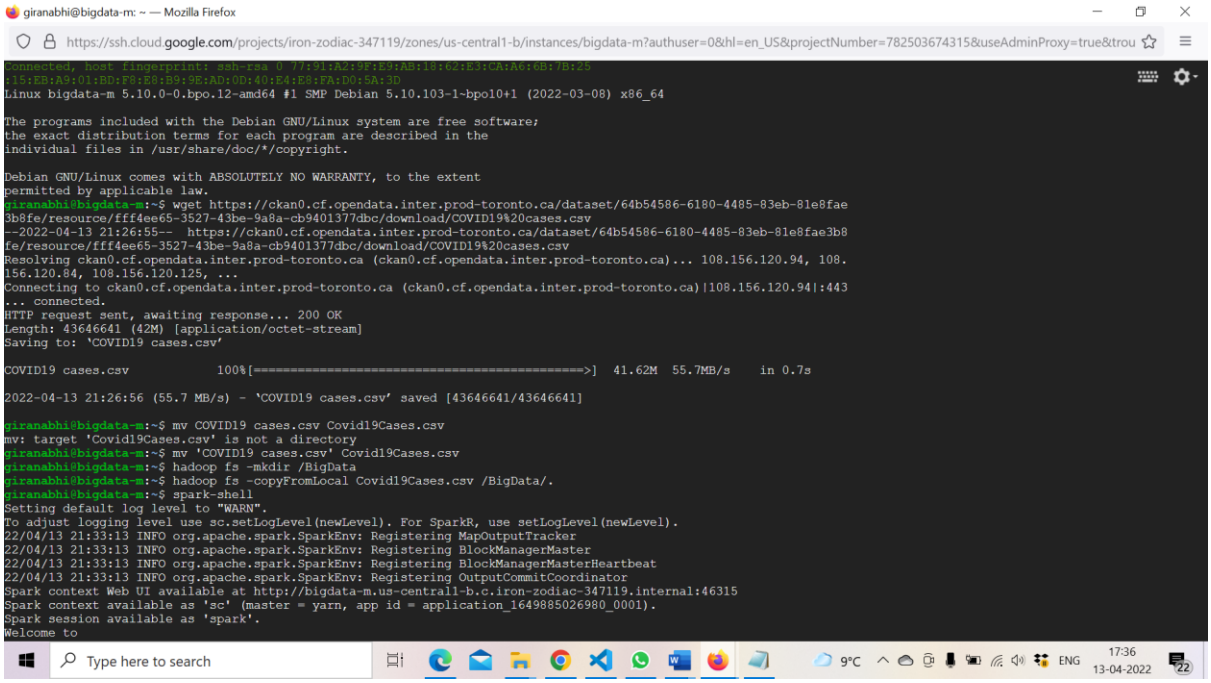
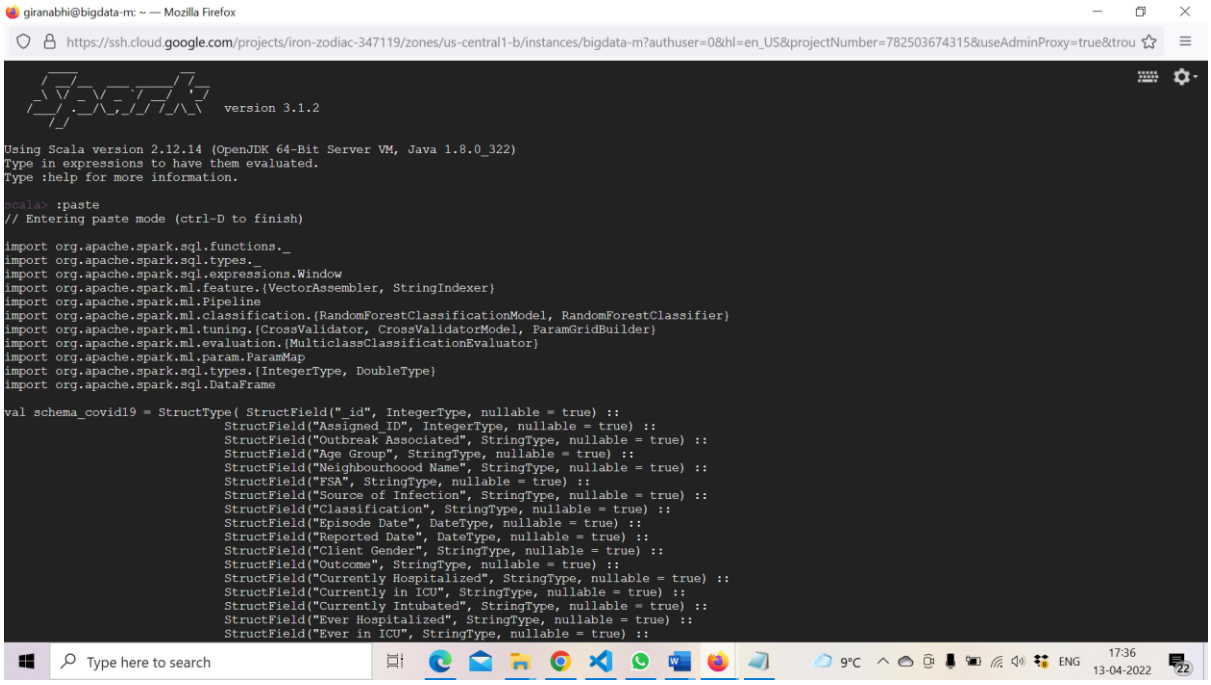
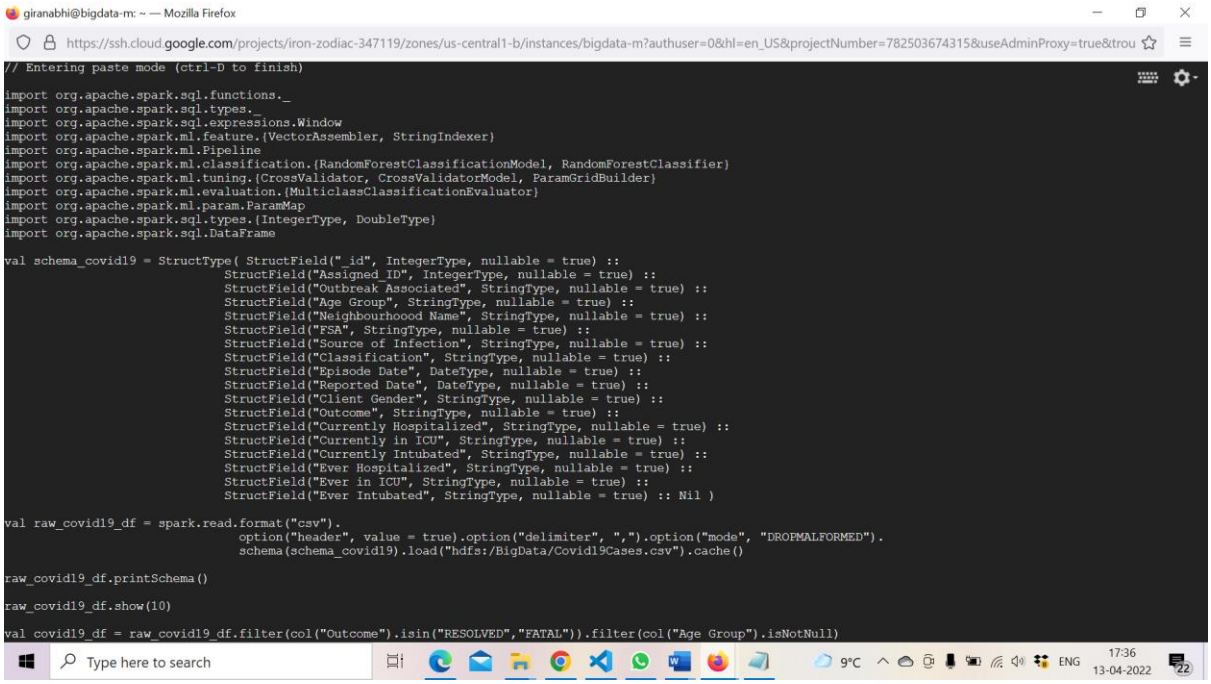
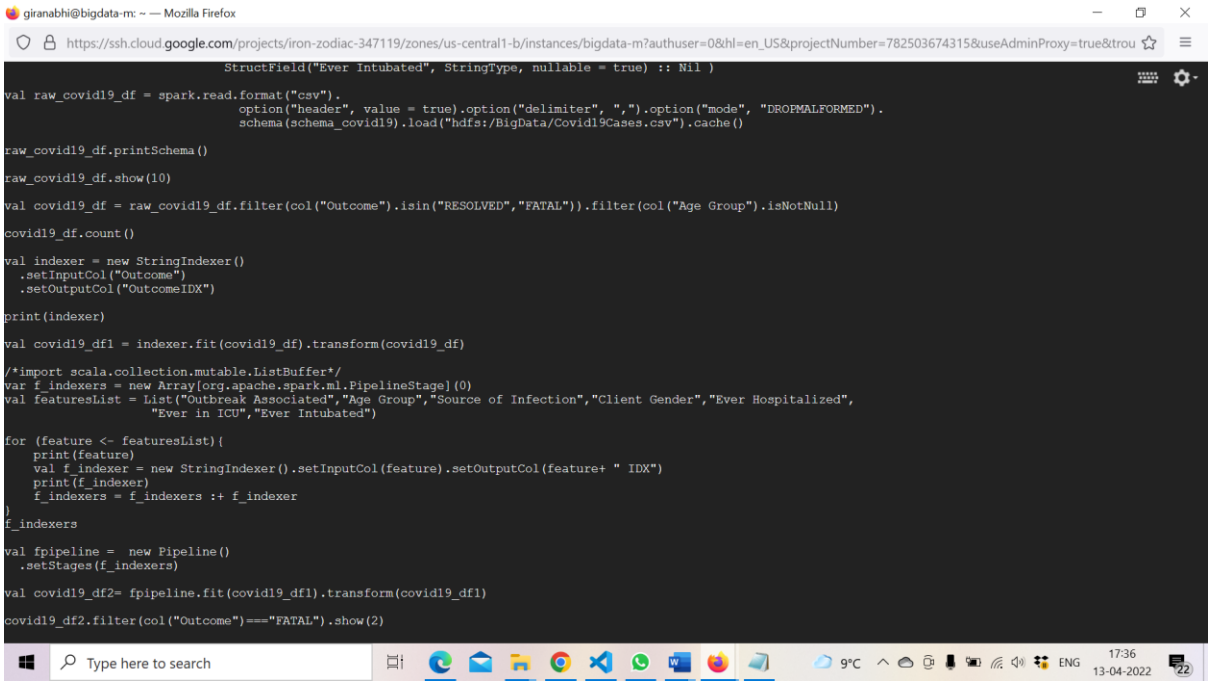
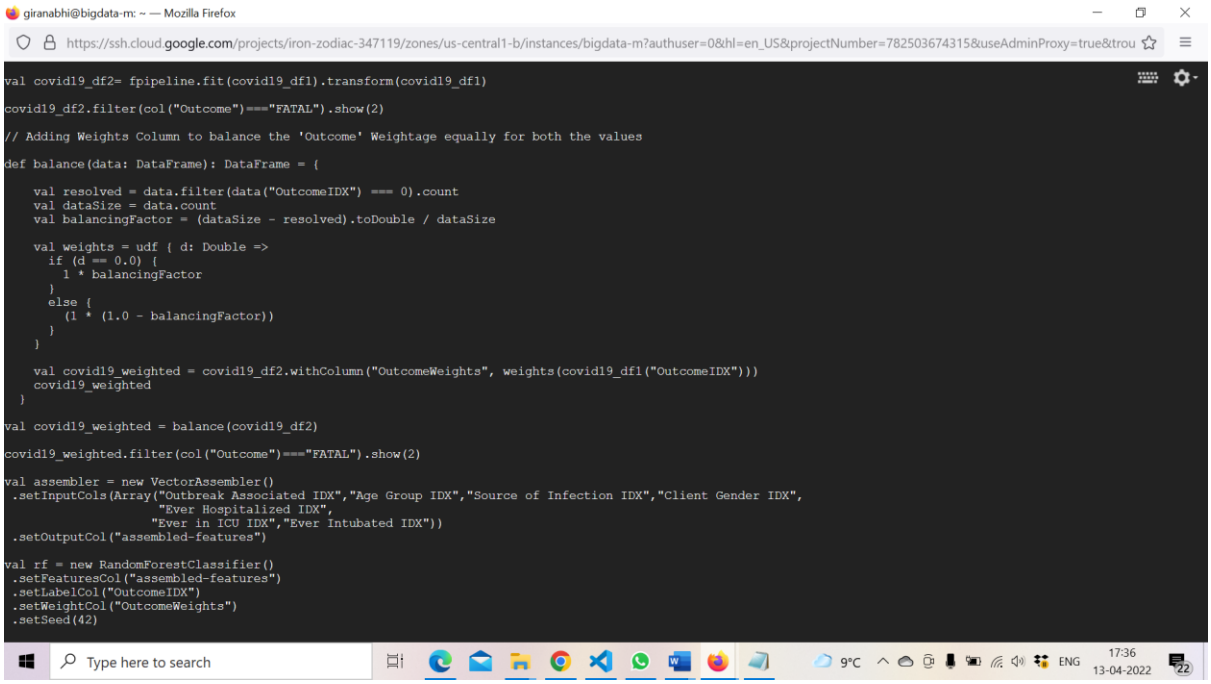
Class weight

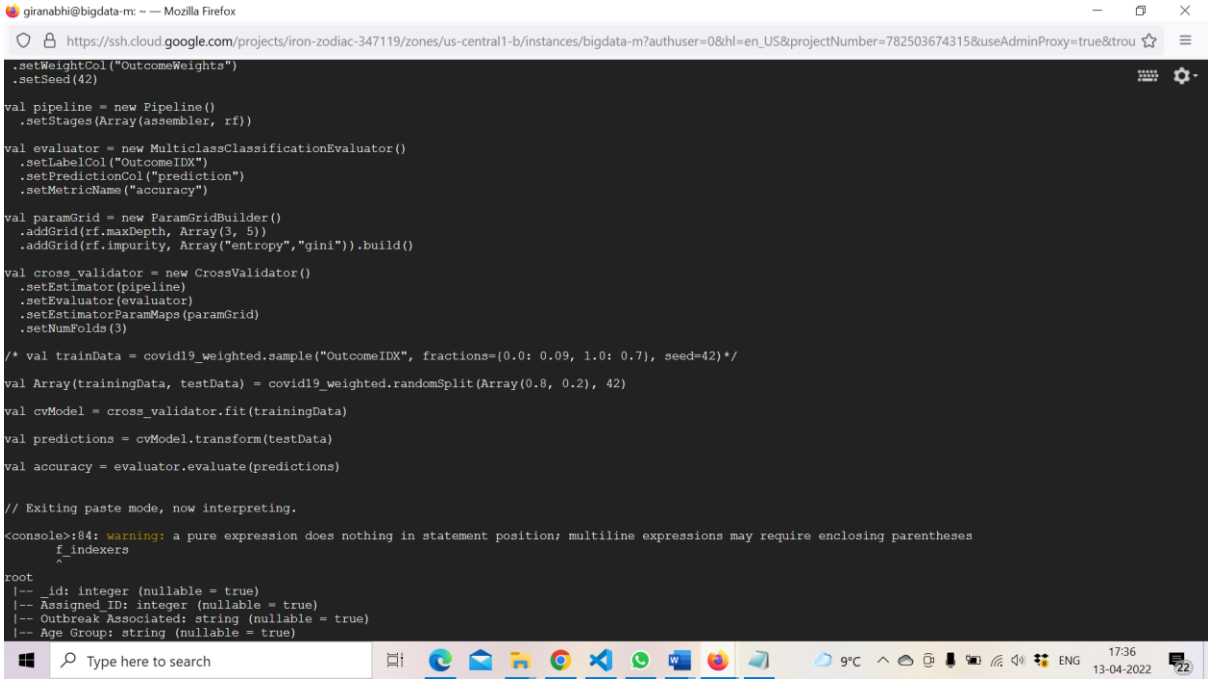


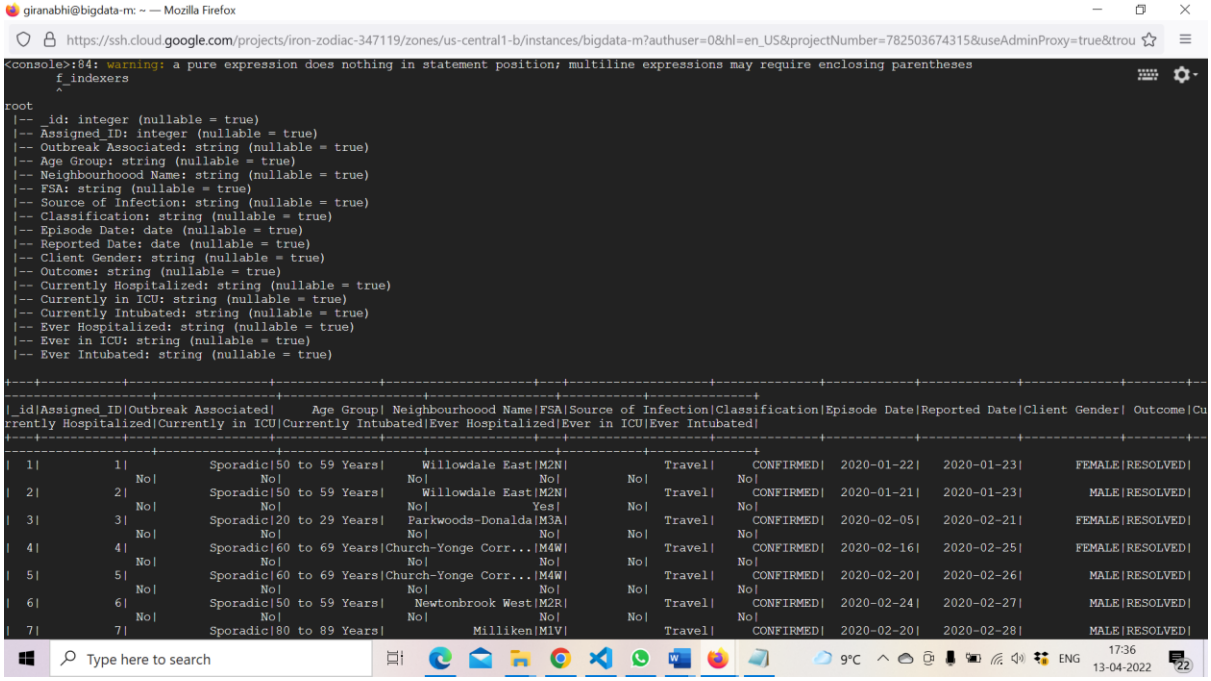


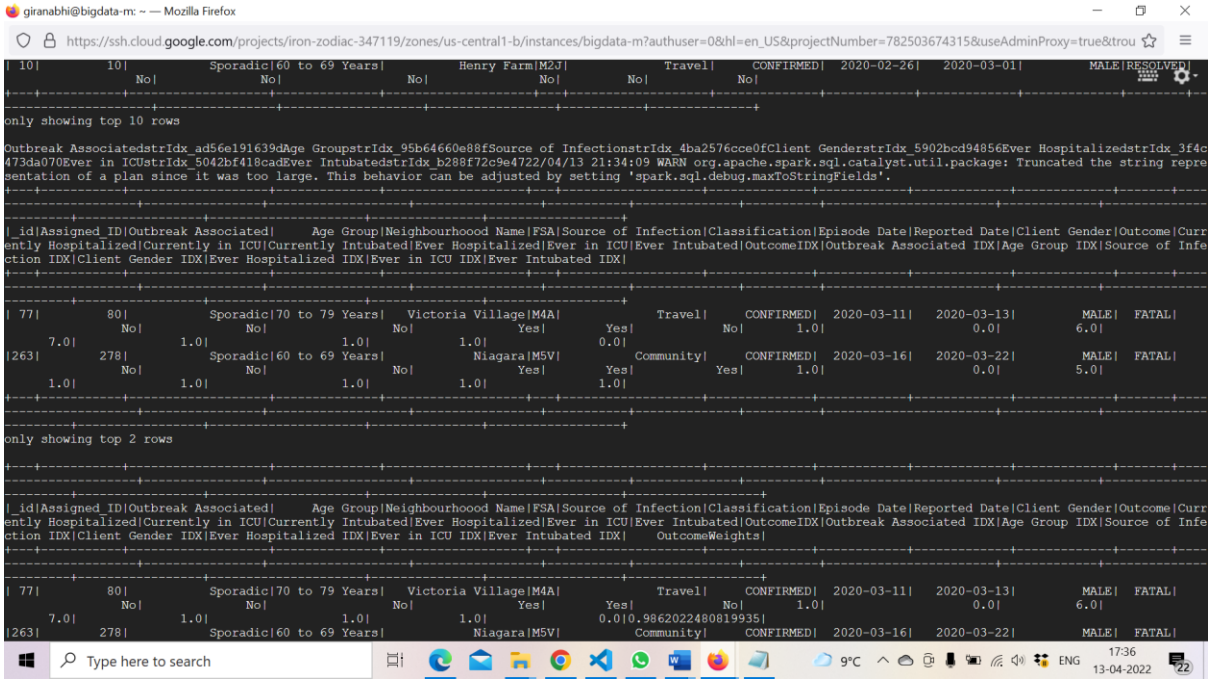

giranabhi@bigdata-m: ~ — Mozilla Firefox
https://ssh.cloud.google.com/projects/iron-zodiac-347119/zones/us-central1-b/instances/bigdata-m?authuser=0&hl=en_US&projectNumber=782503674315&useAdminProxy=true&trou
// Entering paste mode (ctrl-D to finish)
import org.apache.spark.sql.functions._
import org.apache.spark.sql.types._
import org.apache.spark.sql.expressions.Window
import org.apache.spark.ml.feature.{VectorAssembler, StringIndexer}
import org.apache.spark.ml.Pipeline
import org.apache.spark.ml.classification.{RandomForestClassificationModel, RandomForestClassifier}
import org.apache.spark.ml.tuning.{CrossValidator, CrossValidatorModel, ParamGridBuilder}
import org.apache.spark.ml.evaluation.{MulticlassClassificationEvaluator}
import org.apache.spark.ml.param.ParamMap
import org.apache.spark.sql.types.{IntegerType, DoubleType}
import org.apache.spark.sql.DataFrame
val schema_covid19 = StructType(StructField("_id", IntegerType, nullable = true) ::
StructField("Assigned_ID", IntegerType, nullable = true) ::
StructField("Outbreak Associated", StringType, nullable = true) ::
StructField("Age Group", StringType, nullable = true) ::
StructField("Neighbourhoood Name", StringType, nullable = true) ::
StructField("FSA", StringType, nullable = true) ::
StructField("Source of Infection", StringType, nullable = true) ::
StructField("Classification", StringType, nullable = true) ::
StructField("Episode Date", DateType, nullable = true) ::
StructField("Reported Date", DateType, nullable = true) ::
StructField("Client Gender", StringType, nullable = true) ::
StructField("Outcome", StringType, nullable = true) ::
StructField("Currently Hospitalized", StringType, nullable = true) ::
StructField("Currently in ICU", StringType, nullable = true) ::
StructField("Currently Intubated", StringType, nullable = true) ::
StructField("Ever Hospitalized", StringType, nullable = true) ::
StructField("Ever in ICU", StringType, nullable = true) ::
StructField("Ever Intubated", StringType, nullable = true) :: Nil)
val raw_covid19_df = spark.read.format("csv").
option("header", value = true).option("delimiter", ",").option("mode", "DROPMALFORMED").
schema(schema_covid19).load("hdfs:/BigData/Covid19Cases.csv").cache()
raw_covid19_df.printSchema()
raw_covid19_df.show(10)
val covid19_df = raw_covid19_df.filter(col("Outcome").isin("RESOLVED","FATAL")).filter(col("Age Group").isNotNull)
Type here to search
9°C
ENG
17:36
13-04-2022

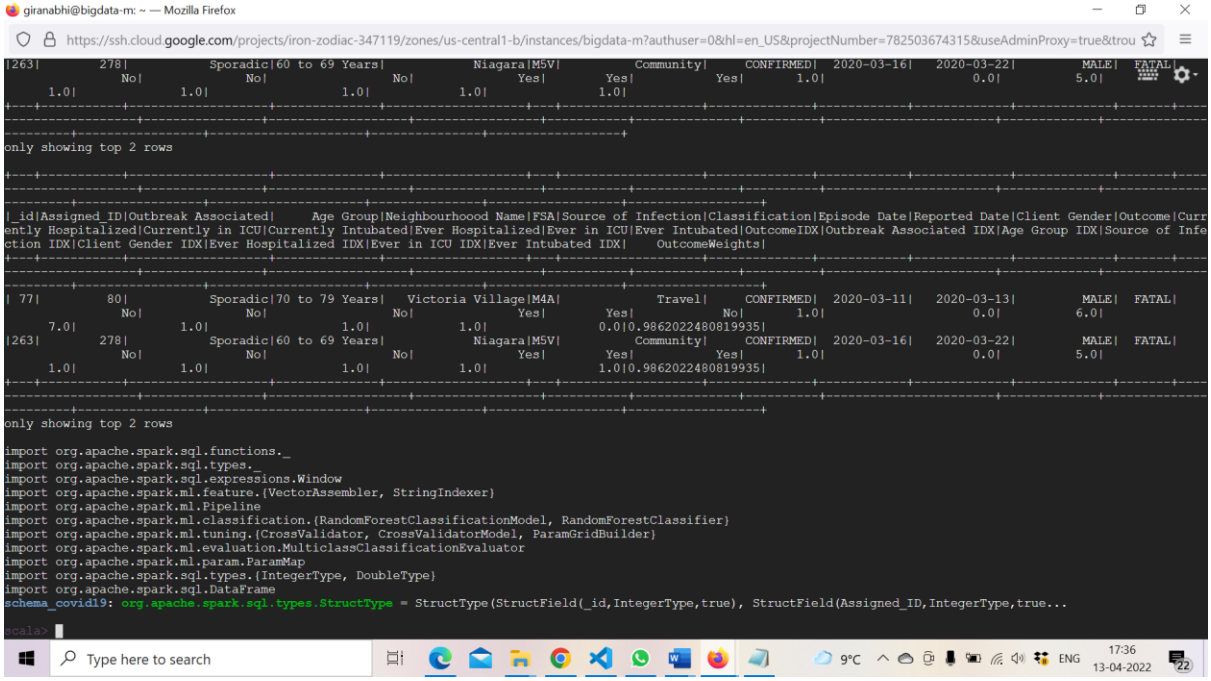

giranabhi@bigdata-m: ~ — Mozilla Firefox
https://ssh.cloud.google.com/projects/iron-zodiac-347119/zones/us-central1-b/instances/bigdata-m?authuser=0&hl=en_US&projectNumber=782503674315&useAdminProxy=true&trou
StructField("Ever Intubated", StringType, nullable = true) :: Nil)
val raw_covid19_df = spark.read.format("csv").
option("header", value = true).option("delimiter", ",").option("mode", "DROPMALFORMED").
schema(schema_covid19).load("hdfs:/BigData/Covid19Cases.csv").cache()
raw_covid19_df.printSchema()
raw_covid19_df.show(10)
val covid19_df = raw_covid19_df.filter(col("Outcome").isin("RESOLVED","FATAL")).filter(col("Age Group").isNotNull)
covid19_df.count()
val indexer = new StringIndexer()
.setInputCol("Outcome")
.setOutputCol("OutcomeIDX")
print(indexer)
val covid19_df1 = indexer.fit(covid19_df).transform(covid19_df)
/*import scala.collection.mutable.ListBuffer*/
var f_indexers = new Array[org.apache.spark.ml.PipelineStage](0)
val featuresList = List("Outbreak Associated","Age Group","Source of Infection","Client Gender","Ever Hospitalized",
"Ever in ICU","Ever Intubated")
for (feature <- featuresList){
print(feature)
val f_indexer = new StringIndexer().setInputCol(feature).setOutputCol(feature+ " IDX")
print(f_indexer)
f_indexers = f_indexers :+ f_indexer
}
f_indexers
val fpipeline = new Pipeline()
.setStages(f_indexers)
val covid19_df2= fpipeline.fit(covid19_df1).transform(covid19_df1)
covid19_df2.filter(col("Outcome")==="FATAL").show(2)
Type here to search
9°C
ENG
17:36
13-04-2022


giranabhi@bigdata-m: ~ — Mozilla Firefox
https://ssh.cloud.google.com/projects/iron-zodiac-347119/zones/us-central1-b/instances/bigdata-m?authuser=0&hl=en_US&projectNumber=782503674315&useAdminProxy=true&trou
val covid19_df2= fpipeline.fit(covid19_df1).transform(covid19_df1)
covid19_df2.filter(col("Outcome")==="FATAL").show(2)
// Adding Weights Column to balance the 'Outcome' Weightage equally for both the values
def balance(data: DataFrame): DataFrame = {
val resolved = data.filter(data("OutcomeIDX") === 0).count
val dataSize = data.count
val balancingFactor = (dataSize - resolved).toDouble / dataSize
val weights = udf { d: Double =>
if (d == 0.0) {
1 * balancingFactor
}
else {
(1 * (1.0 - balancingFactor))
}
}
val covid19_weighted = covid19_df2.withColumn("OutcomeWeights", weights(covid19_df1("OutcomeIDX")))
covid19_weighted
}
val covid19_weighted = balance(covid19_df2)
covid19_weighted.filter(col("Outcome")==="FATAL").show(2)
val assembler = new VectorAssembler()
.setInputCols(Array("Outbreak Associated IDX","Age Group IDX","Source of Infection IDX","Client Gender IDX",
"Ever Hospitalized IDX",
"Ever in ICU IDX","Ever Intubated IDX"))
.setOutputCol("assembled-features")
val rf = new RandomForestClassifier()
.setFeaturesCol("assembled-features")
.setLabelCol("OutcomeIDX")
.setWeightCol("OutcomeWeights")
.setSeed(42)
Type here to search
9°C
ENG
17:36
13-04-2022


giranabhi@bigdata-m: ~ — Mozilla Firefox
https://ssh.cloud.google.com/projects/iron-zodiac-347119/zones/us-central1-b/instances/bigdata-m?authuser=0&hl=en_US&projectNumber=782503674315&useAdminProxy=true&trou
.setWeightCol("OutcomeWeights")
.setSeed(42)
val pipeline = new Pipeline()
.setStages(Array(assembler, rf))
val evaluator = new MulticlassClassificationEvaluator()
.setLabelCol("OutcomeIDX")
.setPredictionCol("prediction")
.setMetricName("accuracy")
val paramGrid = new ParamGridBuilder()
.addGrid(rf.maxDepth, Array(3, 5))
.addGrid(rf.impurity, Array("entropy","gini")).build()
val cross_validator = new CrossValidator()
.setEstimator(pipeline)
.setEvaluator(evaluator)
.setEstimatorParamMaps(paramGrid)
.setNumFolds(3)
/* val trainData = covid19_weighted.sample("OutcomeIDX", fractions={0.0: 0.09, 1.0: 0.7}, seed=42)*/
val Array(trainingData, testData) = covid19_weighted.randomSplit(Array(0.8, 0.2), 42)
val cvModel = cross_validator.fit(trainingData)
val predictions = cvModel.transform(testData)
val accuracy = evaluator.evaluate(predictions)
// Exiting paste mode, now interpreting.
<console>:84: warning: a pure expression does nothing in statement position; multiline expressions may require enclosing parentheses
f_indexers
^
root
|-- _id: integer (nullable = true)
|-- Assigned_ID: integer (nullable = true)
|-- Outbreak Associated: string (nullable = true)
|-- Age Group: string (nullable = true)
Type here to search
9°C
ENG
17:36
13-04-2022


giranabhi@bigdata-m: ~ — Mozilla Firefox
https://ssh.cloud.google.com/projects/iron-zodiac-347119/zones/us-central1-b/instances/bigdata-m?authuser=0&hl=en_US&projectNumber=782503674315&useAdminProxy=true&trou
<console>:84: warning: a pure expression does nothing in statement position; multiline expressions may require enclosing parentheses
f_indexers
root
|-- _id: integer (nullable = true)
|-- Assigned_ID: integer (nullable = true)
|-- Outbreak Associated: string (nullable = true)
|-- Age Group: string (nullable = true)
|-- Neighbourhoood Name: string (nullable = true)
|-- FSA: string (nullable = true)
|-- Source of Infection: string (nullable = true)
|-- Classification: string (nullable = true)
|-- Episode Date: date (nullable = true)
|-- Reported Date: date (nullable = true)
|-- Client Gender: string (nullable = true)
|-- Outcome: string (nullable = true)
|-- Currently Hospitalized: string (nullable = true)
|-- Currently in ICU: string (nullable = true)
|-- Currently Intubated: string (nullable = true)
|-- Ever Hospitalized: string (nullable = true)
|-- Ever in ICU: string (nullable = true)
|-- Ever Intubated: string (nullable = true)
|_id|Assigned_ID|Outbreak Associated| Age Group| Neighbourhoood Name|FSA|Source of Infection|Classification|Episode Date|Reported Date|Client Gender| Outcome|Cu
rrently Hospitalized|Currently in ICU|Currently Intubated|Ever Hospitalized|Ever in ICU|Ever Intubated|
| 1| 1| Sporadic|50 to 59 Years| Willowdale East|M2N| Travel| CONFIRMED| 2020-01-22| 2020-01-23| FEMALE|RESOLVED|
No| No| No| No| No|
| 2| 2| Sporadic|50 to 59 Years| Willowdale East|M2N| Travel| CONFIRMED| 2020-01-21| 2020-01-23| MALE|RESOLVED|
No| No| No| Yes| No|
| 3| 3| Sporadic|20 to 29 Years| Parkwoods-Donalda|M3A| Travel| CONFIRMED| 2020-02-05| 2020-02-21| FEMALE|RESOLVED|
No| No| No| No| No|
| 4| 4| Sporadic|60 to 69 Years|Church-Yonge Corr...|M4W| Travel| CONFIRMED| 2020-02-16| 2020-02-25| FEMALE|RESOLVED|
No| No| No| No| No|
| 5| 5| Sporadic|60 to 69 Years|Church-Yonge Corr...|M4W| Travel| CONFIRMED| 2020-02-20| 2020-02-26| MALE|RESOLVED|
No| No| No| No| No|
| 6| 6| Sporadic|50 to 59 Years| Newtonbrook West|M2R| Travel| CONFIRMED| 2020-02-24| 2020-02-27| MALE|RESOLVED|
No| No| No| No| No|
| 7| 7| Sporadic|80 to 89 Years| Milliken|M1V| Travel| CONFIRMED| 2020-02-20| 2020-02-28| MALE|RESOLVED|
Type here to search
9°C
ENG
17:36
13-04-2022


giranabhi@bigdata-m: ~ — Mozilla Firefox
https://ssh.cloud.google.com/projects/iron-zodiac-347119/zones/us-central1-b/instances/bigdata-m?authuser=0&hl=en_US&projectNumber=782503674315&useAdminProxy=true&trou
| 10| 10| Sporadic|60 to 69 Years| Henry Farm|M2J| Travel| CONFIRMED| 2020-02-26| 2020-03-01| MALE|RESOLVED|
No| No| No| No| No| No|
only showing top 10 rows
Outbreak AssociatedstrIdx_ad56e191639dAge GroupstrIdx_95b64660e88fSource of InfectionstrIdx_4ba2576cce0fClient GenderstrIdx_5902bcd94856Ever HospitalizedstrIdx_3f4c
473da070Ever in ICUstrIdx_5042bf418cadEver IntubatedstrIdx_b288f72c9e4722/04/13 21:34:09 WARN org.apache.spark.sql.catalyst.util.package: Truncated the string repre
sentation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
|_id|Assigned_ID|Outbreak Associated| Age Group|Neighbourhoood Name|FSA|Source of Infection|Classification|Episode Date|Reported Date|Client Gender|Outcome|Curr
ently Hospitalized|Currently in ICU|Currently Intubated|Ever Hospitalized|Ever in ICU|Ever Intubated|OutcomeIDX|Outbreak Associated IDX|Age Group IDX|Source of Infe
ction IDX|Client Gender IDX|Ever Hospitalized IDX|Ever in ICU IDX|Ever Intubated IDX|
| 77| 80| Sporadic|70 to 79 Years| Victoria Village|M4A| Travel| CONFIRMED| 2020-03-11| 2020-03-13| MALE| FATAL|
No| No| No| Yes| Yes| No| 1.0| 0.0| 6.0|
7.0| 1.0| 1.0| 1.0| 0.0|
|263| 278| Sporadic|60 to 69 Years| Niagara|M5V| Community| CONFIRMED| 2020-03-16| 2020-03-22| MALE| FATAL|
No| No| No| Yes| Yes| Yes| 1.0| 0.0| 5.0|
1.0| 1.0| 1.0| 1.0| 1.0|
only showing top 2 rows
|_id|Assigned_ID|Outbreak Associated| Age Group|Neighbourhoood Name|FSA|Source of Infection|Classification|Episode Date|Reported Date|Client Gender|Outcome|Curr
ently Hospitalized|Currently in ICU|Currently Intubated|Ever Hospitalized|Ever in ICU|Ever Intubated|OutcomeIDX|Outbreak Associated IDX|Age Group IDX|Source of Infe
ction IDX|Client Gender IDX|Ever Hospitalized IDX|Ever in ICU IDX|Ever Intubated IDX| OutcomeWeights|
| 77| 80| Sporadic|70 to 79 Years| Victoria Village|M4A| Travel| CONFIRMED| 2020-03-11| 2020-03-13| MALE| FATAL|
No| No| No| Yes| Yes| No| 1.0| 0.0| 6.0|
7.0| 1.0| 1.0| 1.0| 0.0|0.9862022480819935|
|263| 278| Sporadic|60 to 69 Years| Niagara|M5V| Community| CONFIRMED| 2020-03-16| 2020-03-22| MALE| FATAL|
Type here to search
9°C
ENG
17:36
13-04-2022


giranabhi@bigdata-m: ~ — Mozilla Firefox
https://ssh.cloud.google.com/projects/iron-zodiac-347119/zones/us-central1-b/instances/bigdata-m?authuser=0&hl=en_US&projectNumber=782503674315&useAdminProxy=true&trou
|263| 278| Sporadic|60 to 69 Years| Niagara|M5V| Community| CONFIRMED| 2020-03-16| 2020-03-22| MALE| FATAL|
No| No| No| Yes| Yes| Yes| 1.0| 0.0| 5.0|
1.0| 1.0| 1.0| 1.0| 1.0|
only showing top 2 rows
|_id|Assigned_ID|Outbreak Associated| Age Group|Neighbourhoood Name|FSA|Source of Infection|Classification|Episode Date|Reported Date|Client Gender|Outcome|Curr
ently Hospitalized|Currently in ICU|Currently Intubated|Ever Hospitalized|Ever in ICU|Ever Intubated|OutcomeIDX|Outbreak Associated IDX|Age Group IDX|Source of Infe
ction IDX|Client Gender IDX|Ever Hospitalized IDX|Ever in ICU IDX|Ever Intubated IDX| OutcomeWeights|
| 77| 80| Sporadic|70 to 79 Years| Victoria Village|M4A| Travel| CONFIRMED| 2020-03-11| 2020-03-13| MALE| FATAL|
No| No| No| Yes| Yes| No| 1.0| 0.0| 6.0|
7.0| 1.0| 1.0| 1.0| 0.0|0.9862022480819935|
|263| 278| Sporadic|60 to 69 Years| Niagara|M5V| Community| CONFIRMED| 2020-03-16| 2020-03-22| MALE| FATAL|
No| No| No| Yes| Yes| Yes| 1.0| 0.0| 5.0|
1.0| 1.0| 1.0| 1.0| 1.0|0.9862022480819935|
only showing top 2 rows
import org.apache.spark.sql.functions._
import org.apache.spark.sql.types._
import org.apache.spark.sql.expressions.Window
import org.apache.spark.ml.feature.{VectorAssembler, StringIndexer}
import org.apache.spark.ml.Pipeline
import org.apache.spark.ml.classification.{RandomForestClassificationModel, RandomForestClassifier}
import org.apache.spark.ml.tuning.{CrossValidator, CrossValidatorModel, ParamGridBuilder}
import org.apache.spark.ml.evaluation.MulticlassClassificationEvaluator
import org.apache.spark.ml.param.ParamMap
import org.apache.spark.sql.types.{IntegerType, DoubleType}
import org.apache.spark.sql.DataFrame
schema_covid19: org.apache.spark.sql.types.StructType = StructType(StructField(_id,IntegerType,true), StructField(Assigned_ID,IntegerType,true...
scala>
Type here to search
9°C
ENG
17:36
13-04-2022
22