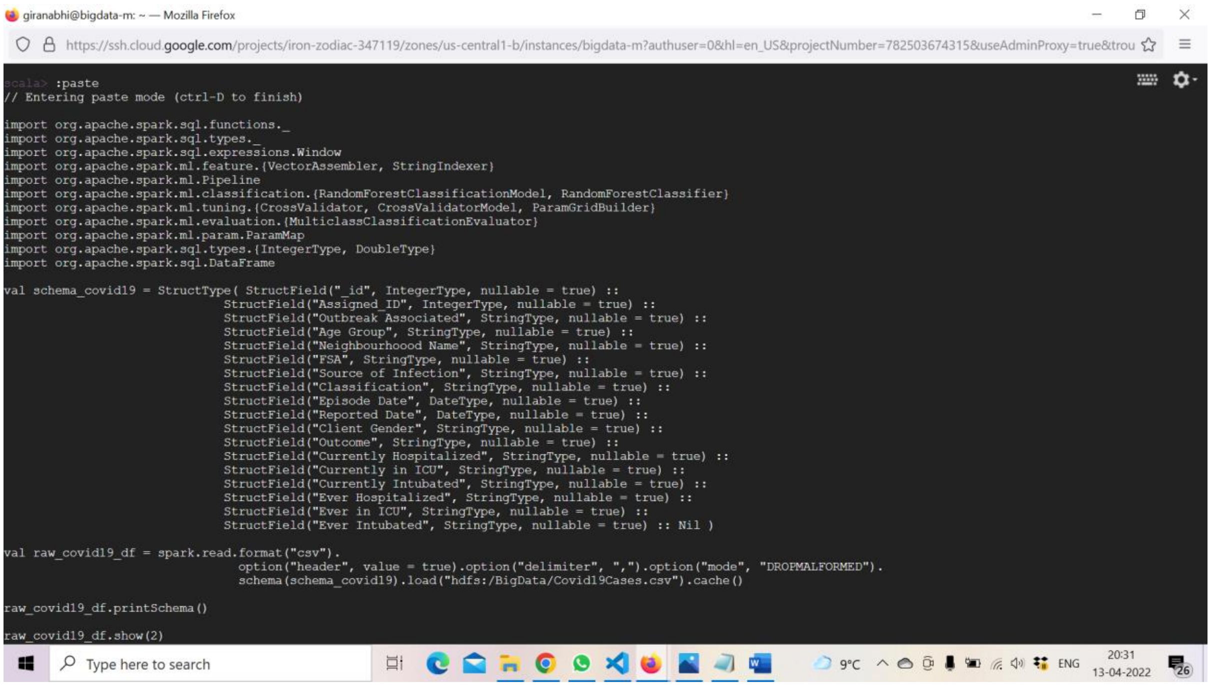
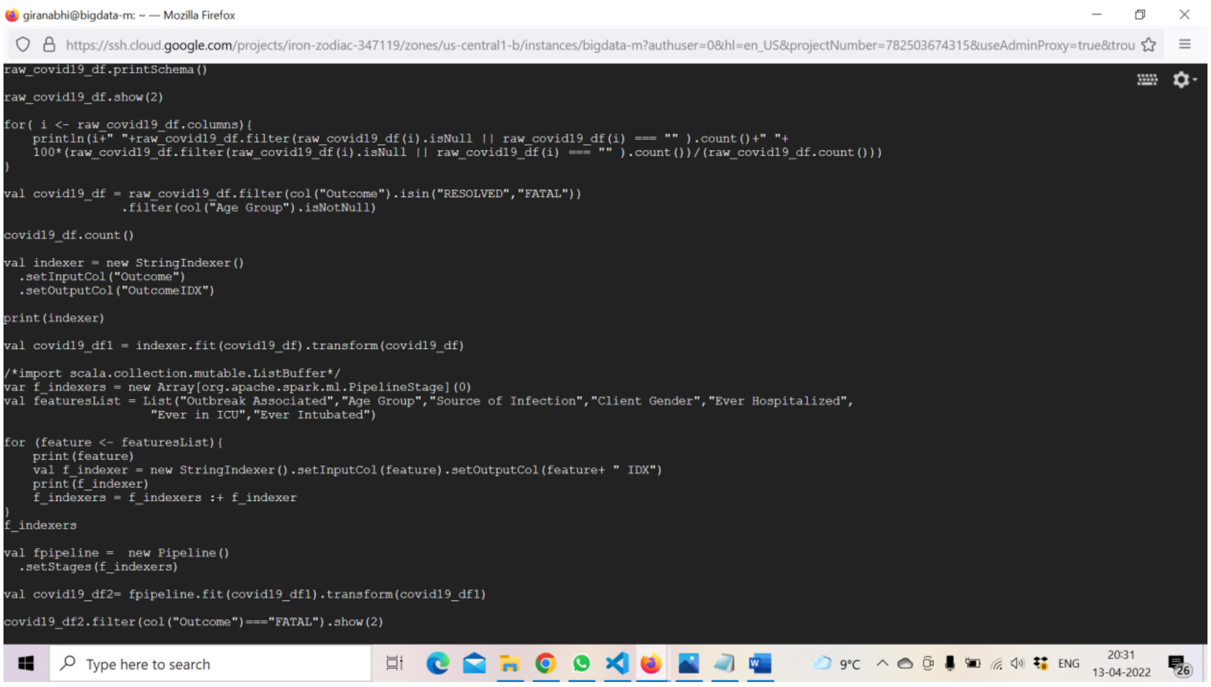
Under Sampling



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

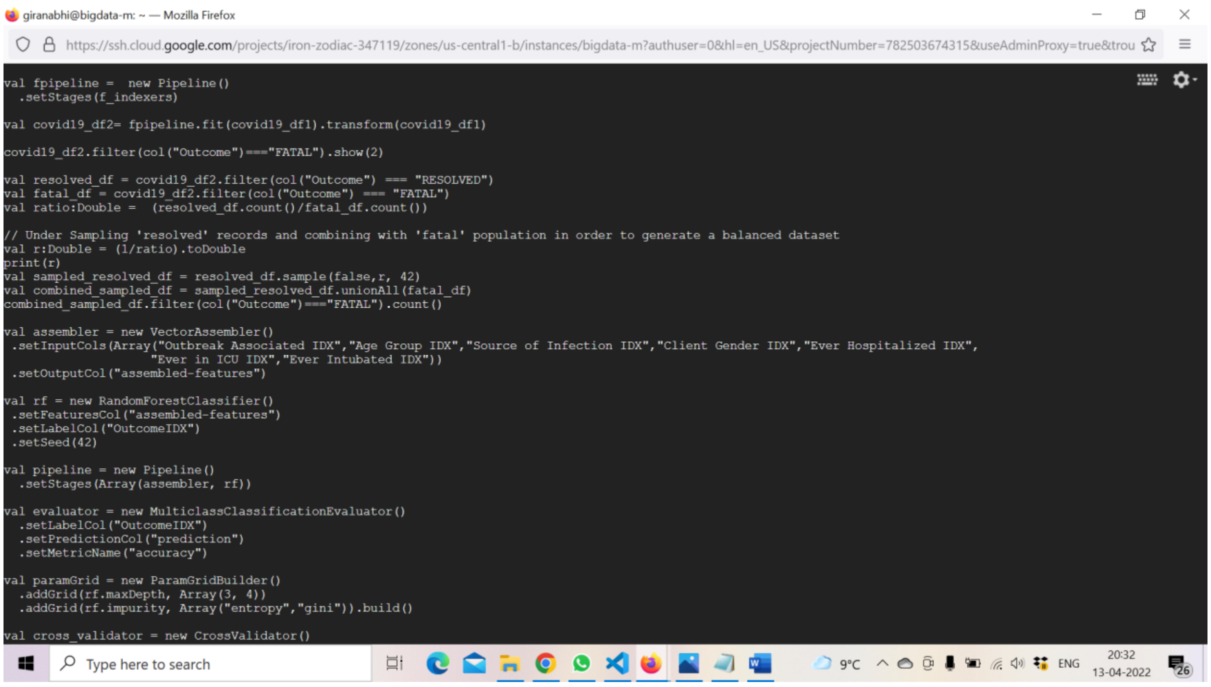
```
scala> :paste
// Entering paste mode (ctrl-D to finish)

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(2)
```



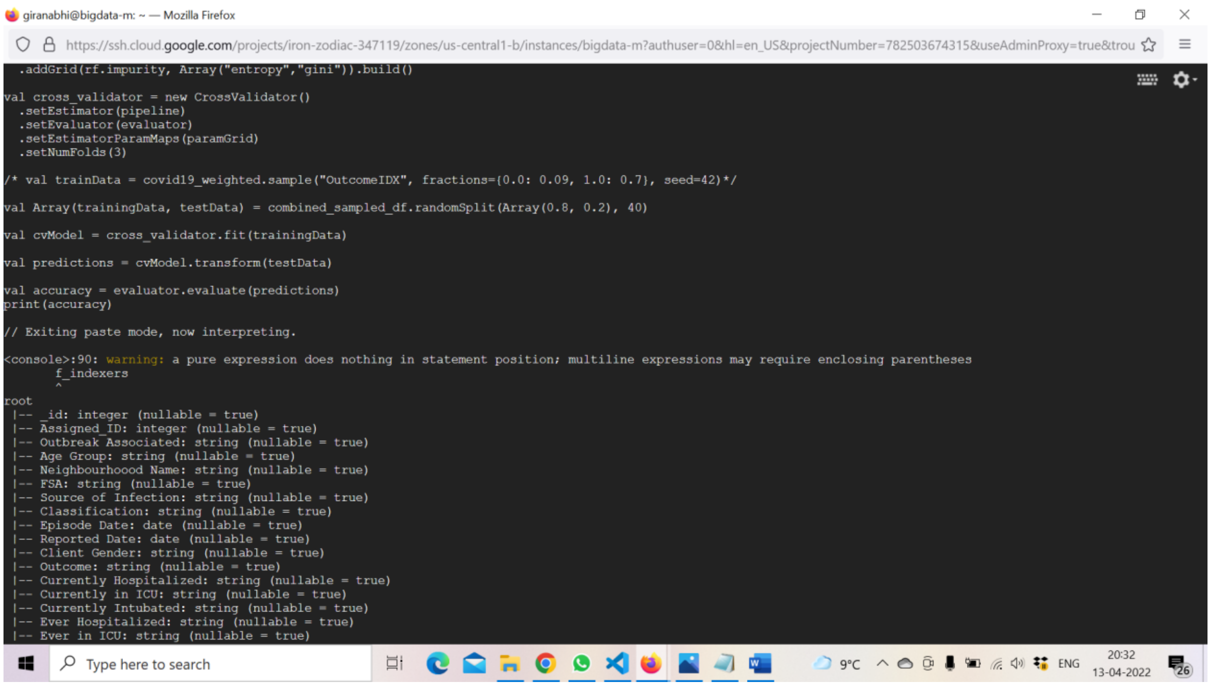
giranabhi@bigdata-m: ~ — Mozilla Firefox

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

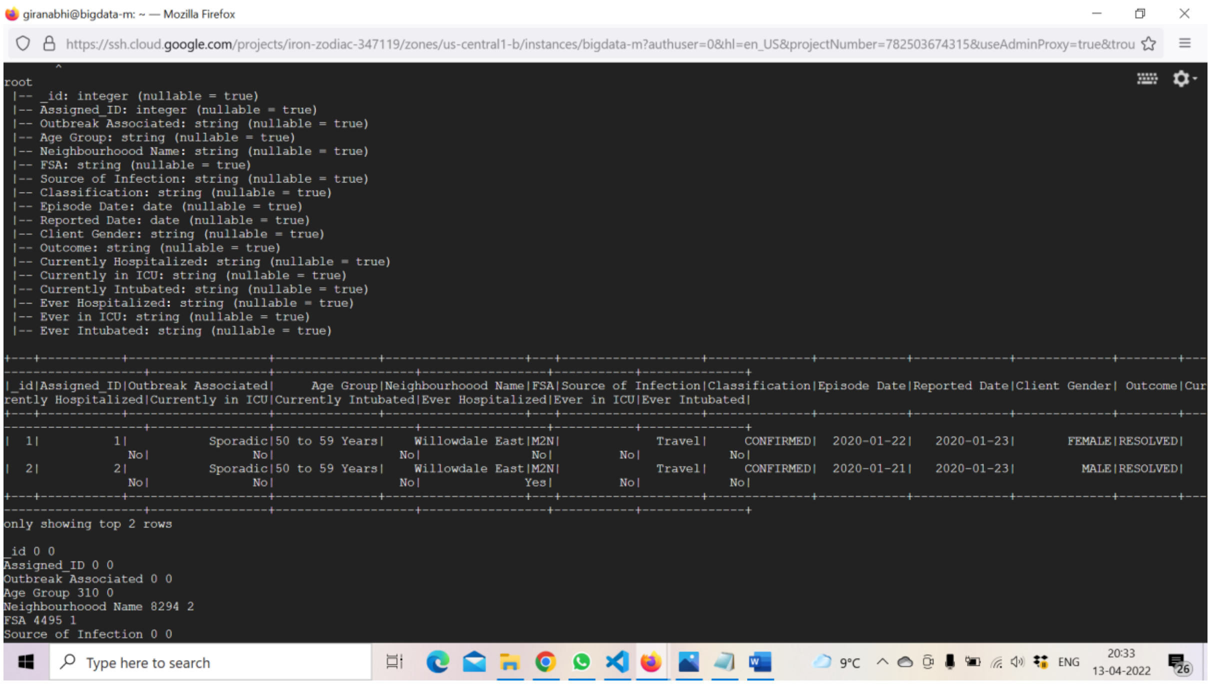
```
raw_covid19_df.printSchema()

raw_covid19_df.show(2)

for( i <- raw_covid19_df.columns){
    println(i+" "+raw_covid19_df.filter(raw_covid19_df(i).isNull || raw_covid19_df(i) === "" ).count()+" "+
    100*(raw_covid19_df.filter(raw_covid19_df(i).isNull || raw_covid19_df(i) === "" ).count())/(raw_covid19_df.count()))
}

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)
```


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
Type here to search
9°C
ENG
20:32
13-04-2022

```
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)

val resolved_df = covid19_df2.filter(col("Outcome") === "RESOLVED")
val fatal_df = covid19_df2.filter(col("Outcome") === "FATAL")
val ratio:Double =  (resolved_df.count()/fatal_df.count())

// Under Sampling 'resolved' records and combining with 'fatal' population in order to generate a balanced dataset
val r:Double = (1/ratio).toDouble
print(r)
val sampled_resolved_df = resolved_df.sample(false,r, 42)
val combined_sampled_df = sampled_resolved_df.unionAll(fatal_df)
combined_sampled_df.filter(col("Outcome")==="FATAL").count()

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")
 .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, 4))
  .addGrid(rf.impurity, Array("entropy","gini")).build()

val cross_validator = new CrossValidator()
```


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
Type here to search
9°C
ENG
20:32
13-04-2022

```
  .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) = combined_sampled_df.randomSplit(Array(0.8, 0.2), 40)

val cvModel = cross_validator.fit(trainingData)

val predictions = cvModel.transform(testData)

val accuracy = evaluator.evaluate(predictions)
print(accuracy)

// Exiting paste mode, now interpreting.

<console>:90: 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)
```

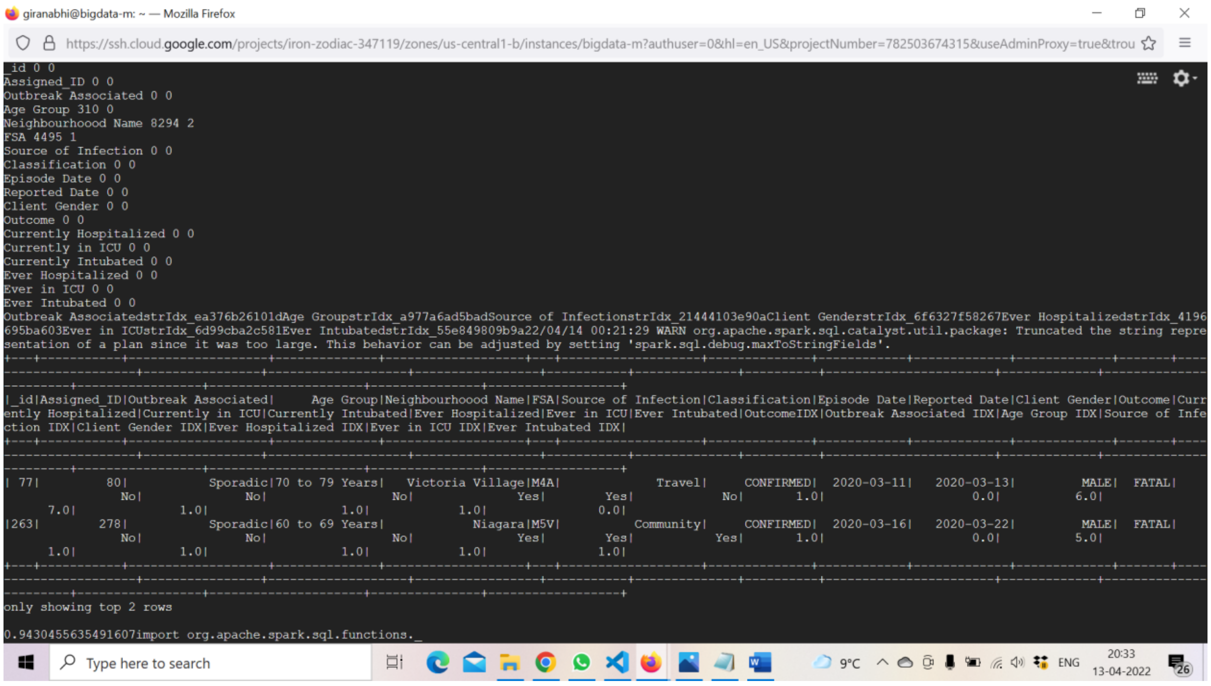


giranabhi@bigdata-m: ~ — Mozilla Firefox

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

```
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|Cur
rently 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|            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|            No|
+---+-----------+-------------------+-------------+-------------------+---+-------------------+--------------+------------+-------------+-------------+--------+---
-------------------+----------------+-------------------+-----------------+-----------+--------------+
only showing top 2 rows

_id 0 0
Assigned_ID 0 0
Outbreak Associated 0 0
Age Group 310 0
Neighbourhoood Name 8294 2
FSA 4495 1
Source of Infection 0 0
```

Type here to search　9°C　ENG　20:33 13-04-2022

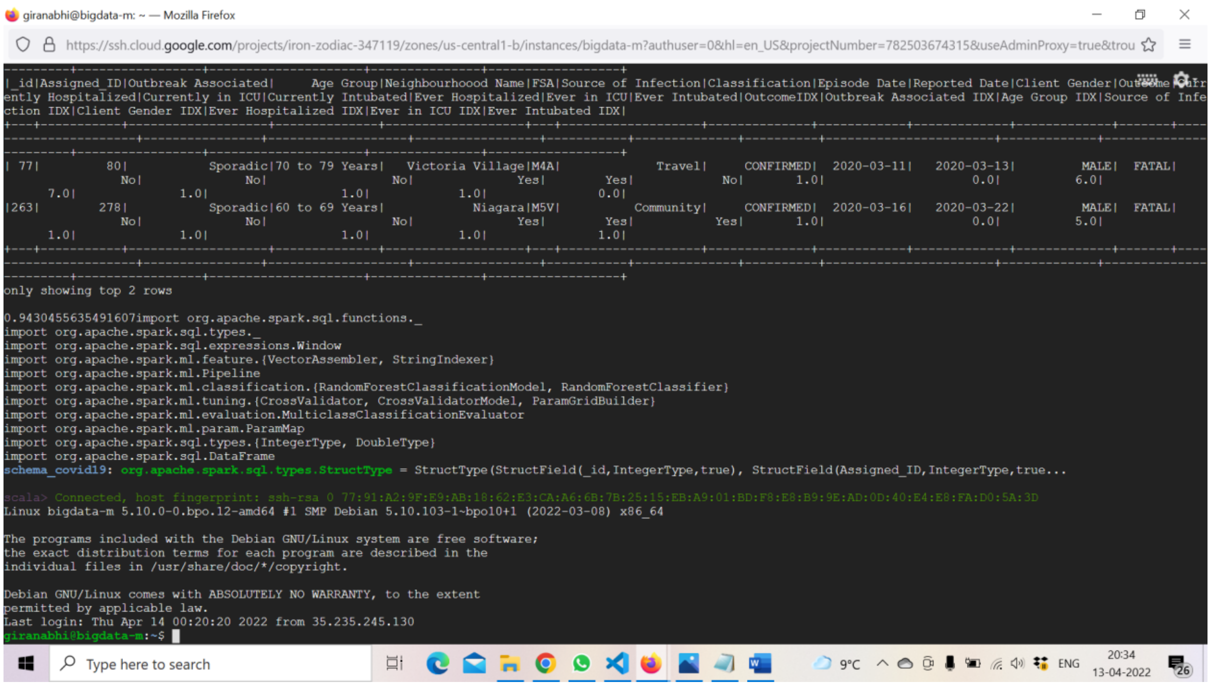


giranabhi@bigdata-m: ~ — Mozilla Firefox

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

```
_id 0 0
Assigned_ID 0 0
Outbreak Associated 0 0
Age Group 310 0
Neighbourhoood Name 8294 2
FSA 4495 1
Source of Infection 0 0
Classification 0 0
Episode Date 0 0
Reported Date 0 0
Client Gender 0 0
Outcome 0 0
Currently Hospitalized 0 0
Currently in ICU 0 0
Currently Intubated 0 0
Ever Hospitalized 0 0
Ever in ICU 0 0
Ever Intubated 0 0
Outbreak AssociatedstrIdx_ea376b26101dAge GroupstrIdx_a977a6ad5badSource of InfectionstrIdx_21444103e90aClient GenderstrIdx_6f6327f58267Ever HospitalizedstrIdx_4196
695ba603Ever in ICUstrIdx_6d99cba2c581Ever IntubatedstrIdx_55e849809b9a22/04/14 00:21:29 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

0.9430455635491607import org.apache.spark.sql.functions._
```

Type here to search　9°C　ENG　20:33 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

```
---------+-----------------+--------------------+---------------+-----------------+
|_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|           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|       Yes|                    1.0|          0.0|           5.0|
      1.0|              1.0|                 1.0|            1.0|              1.0|
+---+-----------+-------------------+--------------+-------------------+---+-------------------+--------------+------------+-------------+-------------+-------+----
------------------+----------------+-------------------+-----------------+-----------+--------------+----------+-----------------------+-------------+--------------
---------+-----------------+--------------------+---------------+-----------------+
only showing top 2 rows

0.9430455635491607import 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> Connected, host fingerprint: ssh-rsa 0 77:91:A2:9F:E9:AB:18:62:E3:CA:A6:6B:7B:25:15:EB:A9:01:BD:F8:E8:B9:9E:AD:0D:40:E4:E8:FA:D0:5A:3D
Linux bigdata-m 5.10.0-0.bpo.12-amd64 #1 SMP Debian 5.10.103-1~bpo10+1 (2022-03-08) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Thu Apr 14 00:20:20 2022 from 35.235.245.130
giranabhi@bigdata-m:~$
```