

```

import org.apache.spark.ml.feature.{Normalizer, PCA, VectorAssembler}
import org.apache.spark.mllib.evaluation.RegressionMetrics
import org.apache.spark.ml.regression._
import org.apache.spark.ml.Pipeline

// Lectura del archivo CSV en un objeto DataFrame
val data = spark.read
  .format("csv")
  .option("delimiter", ";")
  .option("header", "true")
  .option("inferSchema", "true")
  .option("nullValue", "NA")
  .load("hdfs:///user/hdfs/demos/machine_learning/ventas_helados.csv")

// Se construye una tabla temporal con los datos de las ventas
data.createOrReplaceTempView("Ventas")

val DF = data.na.drop

val lista_heladerias = List(3409,3859,3893,5280,6021,6228,7190,7357,9015,9240,11344,12080,12337,12540,13828,14148,14679,15499,15532,15566,15604,16565,16916,17522,17668,17943,18227,18302,21227,18037,20759)

for (heladeria <- lista_heladerias){
  val dataf = DF.filter($"Id_heladeria" === heladeria)

  val split = dataf.randomSplit(Array(0.7, 0.3)) //Se utiliza 70% para entrenamiento y 30% de evaluación
  val training = split(0)
  val test = split(1)

  val assembler = new VectorAssembler()
    .setInputCols(Array("Id_heladeria", "Consumo_In_Situ", "Productos_Premium", "Anyo", "Mes", "Dia", "Tipo", "Temperatura", "HorasAbierto"))
    .setOutputCol("features")

  val gbt = new GBRegressor()
    .setLabelCol("Ventas")
    .setFeaturesCol("features")

  val pipeline = new Pipeline()
    .setStages(Array(assembler, gbt))

  // Generar el modelo para una heladería concreta
  val model = pipeline.fit(training)

  .format("csv")
  .option("delimiter", ";")
  .option("header", "true")
  .option("inferSchema", "true")
  .option("nullValue", "NA")
  .load("hdfs:///user/hdfs/demos/machine_learning/ventas_helados.csv")

// Se construye una tabla temporal con los datos de las ventas
data.createOrReplaceTempView("Ventas")

val DF = data.na.drop

val lista_heladerias = List(3409,3859,3893,5280,6021,6228,7190,7357,9015,9240,11344,12080,12337,12540,13828,14148,14679,15499,15532,15566,15604,16565,16916,17522,17668,17943,18227,18302,21227,18037,20759)

for (heladeria <- lista_heladerias){
  val dataf = DF.filter($"Id_heladeria" === heladeria)

  val split = dataf.randomSplit(Array(0.7, 0.3)) //Se utiliza 70% para entrenamiento y 30% de evaluación
  val training = split(0)
  val test = split(1)

  val assembler = new VectorAssembler()
    .setInputCols(Array("Id_heladeria", "Consumo_In_Situ", "Productos_Premium", "Anyo", "Mes", "Dia", "Tipo", "Temperatura", "HorasAbierto"))
    .setOutputCol("features")

  val gbt = new GBRegressor()
    .setLabelCol("Ventas")
    .setFeaturesCol("features")

  val pipeline = new Pipeline()
    .setStages(Array(assembler, gbt))

  // Generar el modelo para una heladería concreta
  val model = pipeline.fit(training)

  // Almacenar en disco cada modelo construido
  model.write.overwrite.save("RegressionModelHeladeria"+heladeria)

  val predictions = model.transform(test)
  predictions.createOrReplaceTempView("predicciones")

```

```
%spark2.spark

// Lectura del archivo CSV en un objeto DataFrame
val data = spark.read
    .format("csv")
    .option("delimiter", ";")
    .option("header", "true")
    .option("inferSchema", "true")
    .option("nullValue", "NA")
    .load("hdfs:///user/hdfs/demos/machine_learning/ventas_helados.csv")

// Se construye una tabla temporal con los datos de las ventas
data.createOrReplaceTempView("Ventas")
```

Anyo	Mes	Dia	Ventas	Prediccion
2013	5	9	903	1467.3070545683204
2013	5	12	1770	1867.003761537395
2013	5	14	708	1288.7788047880301
2013	5	15	948	1040.8762049213692
2013	5	19	1616	1403.219292983964
2013	5	23	996	1489.9226567412406
2013	5	24	1501	1489.9226567412406
2013	5	27	816	1268.085985849021

Id_heladeria	Des_Heladeria	Comunidad_Autonomas	Consumo_In_Situ	Productos_Premium	HorasAbierto	Fecha	Ano	Mes	Dia
3409	PLAYA POSTIGUET	Comunidad Valenciana	0	0	17	2013-05-09 00:00:00.0	2013	5	9
3409	PLAYA POSTIGUET	Comunidad Valenciana	0	0	17	2013-05-10 00:00:00.0	2013	5	10
3409	PLAYA POSTIGUET	Comunidad Valenciana	0	0	17	2013-05-11 00:00:00.0	2013	5	11
3409	PLAYA POSTIGUET	Comunidad Valenciana	0	0	17	2013-05-12 00:00:00.0	2013	5	12
3409	PLAYA POSTIGUET	Comunidad Valenciana	0	0	17	2013-05-13 00:00:00.0	2013	5	13
3409	PLAYA POSTIGUET	Comunidad Valenciana	0	0	17	2013-05-14 00:00:00.0	2013	5	14
3409	PLAYA POSTIGUET	Comunidad Valenciana	0	0	17	2013-05-15 00:00:00.0	2013	5	15
3409	PLAYA POSTIGUET	Comunidad Valenciana	0	0	17	2013-05-16 00:00:00.0	2013	5	16

Heladeria:

ALICANTE - PLAYA POSTIGUET

Temperatura:

20

Horas abierto:

10

Consumo in situ:

No

Productos premium:

No

Rango de fechas para predicción:

September 28, 2018 -
September 28, 2020

Predicir

Predicción de ventas basada en regresión

