INTRODUÇÃO AO SPARK

//Comandos básicos

// Chamando o spark

spark-shell

// O Spark permite ler comandos a partir de scripts

//Lendo um script externo no scala

:load /home/bitnami/input.scala

// Trabalhando com dataframes

// Importando as bibliotecas necessárias

```
import org.apache.spark.sql.SQLContext
import org.apache.spark.sql.functions._
import java.text.SimpleDateFormat
import java.util.Calendar
import org.apache.spark.sql.Row
import org.apache.spark.sql.types.{StructType, StructField, StringT
ype, IntegerType, FloatType, DoubleType}
import org.apache.spark.ml.feature.VectorAssembler
import org.apache.spark.ml.regression.{LinearRegression, RandomFore
stRegressor}
import org.apache.spark.ml.evaluation.RegressionEvaluator

// Marcando a data e hora do inicio da operação
val starttime = Calendar.getInstance().getTime()
```

// Criando os dados de treinamento a partir de uma sequência de tuplas, e transformando-os em DATAFRAME

```
val dataframeTreinamento = spark.createDataFrame(Seq(
  (10L, 3L, 1L),
  (20L, 6L, 2L),
  (30L, 9L, 3L),
  (50L, 15L, 5L),
  (60L, 18L, 6L),
  (100L, 30L, 10L),
  (120L, 36L, 12L),
  (130L, 39L, 13L),
  (160L, 48L, 16L)
  )).toDF("Coluna_1", "Coluna_2", "label")

//Mostrando o DATAFRAME de treinamento
dataframeTreinamento.show()
```

```
+----+
|Coluna_1|Coluna_2|label|
+-----+
| 10| 3| 1|
```

// Criando os dados de teste a partir de uma sequencia de tuplas, e transformando-os em DATAFRAME

```
val dataframeTeste = spark.createDataFrame(Seq(
    (31L, 6L, 3L),
    (40L, 13L, 4L),
    (82L, 24L, 8L),
    (70L, 23L, 7L),
    (149L, 45L, 15L),
    (140L, 43L, 14L),
    (91L, 27L, 9L)
    )).toDF("Coluna_1", "Coluna_2", "label")

//Mostrando o DATAFRAME de teste
dataframeTeste.show()
```

```
+-----+
|Coluna_1|Coluna_2|label|
+-----+
| 31| 6| 3|
| 40| 13| 4|
| 82| 24| 8|
| 70| 23| 7|
| 149| 45| 15|
| 140| 43| 14|
| 91| 27| 9|
+-----+
```

// Usando comandos SQL com Spark.SQL

//Criando Tabela 1 temporaria para visualizar o dataframeTreinamento usando Spark.SQL

```
dataframeTreinamento.createOrReplaceTempView("minhaTabela1")
```

//Usando comandos SQL para visualizações da tabela 1

```
spark.sql("SELECT * FROM minhaTabela1").show
+----+
|Coluna 1|Coluna 2|label|
+----+
     10 | 3 | 1 |
20 | 6 | 2 |
... | 12 |
130 | 39 | 13 |
160 | 48 | 16 |
     130|
+----+
spark.sql("SELECT Coluna 1 FROM minhaTabela1").show
|Coluna 1|
     10|
     20|
     ...
     130|
    160|
spark.sql("SELECT Coluna_1, Coluna_2 FROM minhaTabela1 WHERE Coluna
1 > 100").show
+----+
|Coluna_1|Coluna_2|
+----+
   120 | 36 |
130 | 39 |
160 | 48 |
```

// Usando o vector assembler

// VectorAssembler() para combinar 2 ou mais colunas em um único vetor de features.

```
//Instanciando um assembler para criar um único vetor de features
val assembler = new VectorAssembler().setInputCols(Array( "Coluna 1
", "Coluna 2")).setOutputCol("features")
//Aplicando a transformacao com o assembler nos dados de treinament
val dataframeTreinamento2 = assembler.transform(dataframeTreinament
0)
//Mostrando os dados e esquema
dataframeTreinamento2.show()
+----+
|Coluna_1|Coluna_2|label|
                          features|
+----+
      10| 3| 1| [10.0,3.0]|
20| 6| 2| [20.0,6.0]|
      20|
    ...| ...| ...|
130| 39| 13|[
160| 48| 16|[
              39| 13|[130.0,39.0]|
48| 16|[160.0,48.0]|
     ---+----+
dataframeTreinamento2.printSchema()
// O resultado deve ser como abaixo
|-- Coluna 1: long (nullable = false)
|-- Coluna_2: long (nullable = false)
|-- label: long (nullable = false)
|-- features: vector (nullable = true)
//aplicando a transformacao com o assembler nos dados de teste
val dataframeTeste2 = assembler.transform(dataframeTeste)
```

// Aplicando regressão linear nos dados

// Criando um modelo de regressão linear

```
// Instanciando o modelo LR
val modelo = new LinearRegression().setLabelCol("label").setFeature
sCol("features")

//Treinando o modelo com os dados de treinamento
val modeloTreinado = modelo.fit(dataframeTreinamento2)

// Fazendo Predicoes com os dados de teste ou desconhecidas
val fazendoPredicoes = modeloTreinado.transform(dataframeTeste2)

// Selecionando as colunas para mostrar os resultados
fazendoPredicoes.select("prediction", "label", "features").show()
```

```
// Criando avaliadores para as metricas do modelo
val avaliador = new RegressionEvaluator().setLabelCol("label").setP
redictionCol("prediction").setMetricName("rmse")
val Test_rmse = avaliador.evaluate(fazendoPredicoes)

val avaliador2 = new RegressionEvaluator().setLabelCol("label").set
PredictionCol("prediction").setMetricName("r2")
val Test_R2 = avaliador2.evaluate(fazendoPredicoes)

//Imprimindo resultados na tela
println("A RAIZ DO ERRO MÉDIO QUADRÁTICO (RMSE) nos dados de teste
é = " + Test_rmse)
println("O R2 nos dados de teste é: " + Test_R2);
```

```
A RAIZ DO ERRO MÉDIO QUADRÁTICO (RMSE) nos dados de teste \acute{e}=0.48 79500364742662 O R2 nos dados de teste \acute{e}:0.98674242424243
```

```
val endtime = Calendar.getInstance().getTime()
val elapsedtime = ((endtime.getTime() - starttime.getTime())/1000)
.toString;

//Imprimindo tempo total na tela
println("Tempo total da operação = " + elapsedtime + " segundos.");
```

// Obtendo alguns parâmetros do modelo

// O intercepto é o valor do eixo Y quando X=0.

modeloTreinado.intercept

```
res11: Double = 3.629261699187618E-15
```

// Demais parâmetros do modelo, coeficientes 'a' e 'b'

modeloTreinado.coefficients

res12: org.apache.spark.ml.linalg.Vector = [-6.175859152393819E-17, 0.33333333333333333]

// Exemplo de Machine Learning nos dados do Prouni

// Importando as bibliotecas necessárias

```
import org.apache.spark.sql.SQLContext
import org.apache.spark.sql.functions._
import java.text.SimpleDateFormat
import org.apache.spark.sql.Row
import org.apache.spark.sql.types.{StructType, StructField, StringT
ype, IntegerType, FloatType, DoubleType}
import org.apache.spark.ml.feature.VectorAssembler
import org.apache.spark.ml.regression.{LinearRegression, RandomFore
stRegressor, GBTRegressor}
import org.apache.spark.ml.evaluation.RegressionEvaluator
import org.apache.spark.ml.attribute.Attribute
import org.apache.spark.ml.attribute.Attribute
```

// Esquema da tabela do nosso arquivo

```
val customSchema = StructType(Array(
       StructField("uf busca", StringType, true),
       StructField("cidade busca", StringType, true),
       StructField("universidade nome", StringType, true),
       StructField("campus nome", StringType, true),
       StructField("nome", StringType, true),
       StructField("grau", StringType, true),
       StructField("turno", StringType, true),
       StructField("mensalidade", DoubleType, true),
       StructField("bolsa_integral_cotas", IntegerType, true),
       StructField("bolsa integral ampla", IntegerType, true),
       StructField("bolsa parcial cotas", IntegerType, true),
       StructField("bolsa parcial ampla", IntegerType, true),
       StructField("nota integral ampla", DoubleType, true),
       StructField("nota integral cotas", DoubleType, true),
       StructField("nota parcial ampla", DoubleType, true),
       StructField("nota parcial cotas", DoubleType, true)))
```

```
val df = spark.read.format("csv").option("delimiter",",").option("q
uote","").option("header", "true").schema(customSchema).load("curso
s-prouni.csv")
```

// Verificando o conteúdo lido do arquivo

```
scala> df.show(5)
```

uf_busca cidade_busca	universidade_nome	campus_nome	nome	grau	turno m	ensalidade bolsa_inte
AC Cruzeiro do Sul AC Cruzeiro do Sul AC Cruzeiro do Sul	Universidade Paul Centro Università Faculdade Educaci Universidade Paul Centro Università	PAP CRUZEIRO DO S AC - CRUZEIRO DO CRUZEIRO DO SUL	Administração Administração Administração	Bacharelado Cur Bacharelado Cur Bacharelado Cur	so a Distância so a Distância so a Distância	289.0 298.0 325.0 319.0 298.0

// Verificando o conteúdo lido do arquivo

```
df.printSchema()
```

```
root
 |-- uf busca: string (nullable = true)
 |-- cidade_busca: string (nullable = true)
 |-- universidade nome: string (nullable = true)
 |-- campus nome: string (nullable = true)
 |-- nome: string (nullable = true)
 |-- grau: string (nullable = true)
 |-- turno: string (nullable = true)
 |-- mensalidade: double (nullable = true)
 |-- bolsa integral cotas: integer (nullable = true)
 |-- bolsa integral ampla: integer (nullable = true)
 |-- bolsa parcial cotas: integer (nullable = true)
 |-- bolsa parcial ampla: integer (nullable = true)
 |-- nota_integral_ampla: double (nullable = true)
 |-- nota integral cotas: double (nullable = true)
 |-- nota_parcial_ampla: double (nullable = true)
 |-- nota parcial cotas: double (nullable = true)
```

// Realizando consultas nos dados

```
scala> df.select("uf_busca").distinct().count()
res2: Long = 27

scala> df.select("cidade_busca").distinct().count()
res3: Long = 1158
```

```
scala> df.select("universidade nome").distinct().count()
res4: Long = 1298
scala> df.select("campus nome").distinct().count()
res5: Long = 4770
scala> df.select("nome").distinct().count()
res6: Long = 338
scala> df.select("grau").distinct().show()
+----+
                grau
+----+
             Direito
|Gestão de Recurso...|
|Engenharia de Pro...|
|Design de Interiores|
   Engenharia Civil|
| Gestão da Qualidade|
|Estética e Cosmética|
| Pesquisa e Gestã...|
|Letras - Portuguê...|
        Agronegócio|
|Arquitetura e Urb...|
|Manutenção de Aer...|
         Enfermagem |
|Letras - Portuguê...|
             Física|
            Teologia|
     Banco de Dados|
|Análise e Desenvo...|
| 333 - CENTRO - P...|
          Agronomia|
+----+
only showing top 20 rows
scala> df.select("turno").distinct().show()
```

```
----+
              turno|
| FACULDADE SÃO FID...|
|Engenharia de Pro...|
           Medicina|
|Letras - Portuguê...|
         Odontologia|
         Enfermagem|
           Matutino|
         Matemática|
       KM 102 (101|
  Curso a Distância|
         Vespertino|
| 60 (NOVA SEDE) -...|
|UNIDADE SEDE - Mo...|
| Ciências Biológicas|
          Psicologia|
      Administração|
          Pedagogia|
        Bacharelado|
       Licenciatura|
| 1171 - CENTRO - ...|
+----+
only showing top 20 rows
```

// Indexação

// Instanciando o Indexer

```
val indexer = new StringIndexer().setInputCol("uf_busca").setOutput
Col("uf_buscaIndex").setHandleInvalid("skip").fit(df)
```

// Aplicando a transformação no Dataframe

```
val indexed = indexer.transform(df)
```

// Verificando o Dataframe com a coluna indexada

indexed.show()

minter it is a market and the second of the										
need: graul turneles	matimerialse,	ntegral_cetas(belse_in	togral_empGothelse_p	arctel_cotas!bolia_g	ertist, esplainsta, s	stogrei_empleinote_is	tograf, catasireta, j	parcial,ample/reta_p	ectal_ontolef_b	vare[index
stracinizacionelado/curos a sistância)	289.80	41	4	(will)	8411	173.71	100.00	6031	mil.	28.8
stracke/EachereSalo/Curse a distancial	250.00		9611	relli	96717	646.34	m/31	9633	6031	25-8
streckolkecherelekolloren a Sistèrica-	125.00	76(1)	96013	31	70.111	P6131	eo(1)	877.620	eo31.	4.61
utração mectarolado curos a mistância	310.0	11	1602	rails	8641.)	400.00	91(21)	0.01	91/31	39.8
strackol/sachareGelosCurse a Sistânxia	256-91	6621	MG1	- 11	M-511	Pa [1]	6931	548.00	PV01	35.8
atmic Solikiethers(Seto) tecturns:	915-015	4001	16,011							

// Verificando apenas as duas colunas envolvidas na transação

```
indexed.select("uf_busca", "uf_buscaIndex").distinct().show()
+----+
|uf_busca|uf_buscaIndex|
                  8.0|
      PA|
      AL|
                22.0|
                  1.0|
      MG |
                 10.0|
      DF|
      TOI
                  20.0|
                 15.0|
      MA|
                  3.01
      RS|
      RJ|
                  6.0|
      BA|
                  4.0|
                 25.0|
                 19.0|
      PI|
                 17.0|
      RO|
                  2.0|
      PR|
      | TM
                  11.0|
      RR |
                  26.0|
      MS|
                 14.0|
                 23.0|
      SEI
      RN|
                 21.0|
      SCI
                  5.0|
      CE |
                 13.0|
only showing top 20 rows
```

```
val indexer2 = new StringIndexer().setInputCol("cidade busca").setO
utputCol("cidade buscaIndex").fit(indexed)
val indexed2 = indexer2.transform(indexed)
val indexer3 = new StringIndexer().setInputCol("universidade nome")
.setOutputCol("universidade nomeIndex").fit(indexed2)
val indexed3 = indexer3.transform(indexed2)
val indexer4 = new StringIndexer().setInputCol("campus nome").setOu
tputCol("campus nomeIndex").fit(indexed3)
val indexed4 = indexer4.transform(indexed3)
val indexer5 = new StringIndexer().setInputCol("nome").setOutputCol
("nomeIndex").fit(indexed4)
val indexed5 = indexer5.transform(indexed4)
val indexer6 = new StringIndexer().setInputCol("grau").setOutputCol
("grauIndex").fit(indexed5)
val indexed6 = indexer6.transform(indexed5)
val indexer7 = new StringIndexer().setInputCol("turno").setOutputCo
l("turnoIndex").fit(indexed6)
val indexed7 = indexer7.transform(indexed6)
```

// Verificar o resultado da transformmação

```
indexed7.show()
```

// Listando apenas as colunas indexadas.

indexed7.select("uf_busca", "uf_buscaIndex", "universidade_nomeInde
x", "nomeIndex", "grauIndex", "turnoIndex").distinct().show()

		_	nomeIndex +		ıf_buscaIndex univ	busca uf __
·			0.0		18.0	AM
0.0	0.0		0.0	2.0	15.0	MA
1.0	0.0		0.0	195.0	1.0	MG
0.0	0.0		0.0	2.0	2.0	PR
0.0	0.0		0.0	5.0	6.0	RJ
1.0	0.0			64.0	3.0	RS
1.0	0.0		0.0	421.0	0.0	SP
2.0	0.0		0.0	231.0	0.0	SP
3.0	0.0		67.0	214.0	3.0	RS
1.0	1.0		9.0	599.0	3.0	RS
2.0	0.0		25.0	61.0	9.0	PE
1.0	0.0		25.0	425.0	5.0	SC
0.0	2.0		28.0	0.0	10.0	DF
0.0	1.0		87.0	11.0	0.0	SP
1.0	0.0		33.0	760.0	2.0	PR
1.0	0.0		33.0	123.0	3.0	RS
0.0	2.0		31.0	16.0	12.0	ES
1.0	2.0		31.0	174.0	0.0	SP
2.0	0.0		2.0	75.0	10.0	DF
2.0	0.0		2.0	1099.0	12.0	ES
+	+-	+	+		+	+

// Criando as bases de treinamento e teste

//Instanciando um assembler para criar um único vetor de features

```
val assembler = new VectorAssembler().setInputCols(Array( "uf_busca
Index", "nomeIndex", "grauIndex", "turnoIndex")).setOutputCol("feat
ures")
```

//Aplicando a transformacao com o assembler nos dados de treinamento

```
val indexed8 = assembler.transform(indexed7)
```

// Split do dataframe indexado

```
val splitDF = indexed8.randomSplit(Array(1,1))
```

//Aplicando o split para criar dataframes de treino e teste

```
val (dataframeTreinamento, dataframeTeste) = (splitDF(0), splitDF(1))
```

// Aprendizado de máquina com os dados

//Instanciando o modelo RandomForestRegressor

```
val modelo = new RandomForestRegressor().setLabelCol("mensalidade")
.setFeaturesCol("features").setMaxBins(338).setNumTrees(100).setSee
d(1)
```

//Treinando o modelo com os dados de treinamento

```
val modeloTreinado = modelo.fit(indexed8)
```

//Fazendo Predicoes com os dados de teste ou desconhecidas

```
val fazendoPredicoes = modeloTreinado.transform(dataframeTeste)
```

```
fazendoPredicoes.select("uf_busca", "nome", "grau", "turno", "prediction", "mensalidade").show()
```

// Criando avaliadores para as métricas do modelo, usando o RMSE e o R2

```
val avaliador = new RegressionEvaluator().setLabelCol("mensalidade"
).setPredictionCol("prediction").setMetricName("rmse")

val Test_rmse = avaliador.evaluate(fazendoPredicoes)

val avaliador2 = new RegressionEvaluator().setLabelCol("mensalidade").setPredictionCol("prediction").setMetricName("r2")

val Test_R2 = avaliador2.evaluate(fazendoPredicoes)

println("A RAIZ DO ERRO MÉDIO QUADRATICO (RMSE) nos dados de teste = " + Test_rmse)

println("O R2 nos dados de teste : " + Test_R2);
```

//Criando uma tabela temporaria para fazer queries nos resultados usando SQL

fazendoPredicoes.createOrReplaceTempView("minhaTabela")

spark.sql("SELECT uf_busca, grau, turno, nome, prediction, mensalid
ade FROM minhaTabela WHERE uf_busca = 'RJ'").show

						nome		
 						Ciências Contábeis		
	RJ	Tecnológico	Curso	а	Distância	Processos Gerenciais	304.01096607823195	325.0
	RJ	Bacharelado	Curso	а	Distância	Administração	333.34754851326505	279.0
	RJ	Bacharelado	Curso	а	Distância	Enfermagem	595.8265423619935	519.0
	RJ	Bacharelado			Noturno	Administração	979.6470106030273	720.0
	RJ	Bacharelado			Noturno	Biomedicina	1220.7292393503467	920.0
	RJ	Bacharelado			Noturno	Direito	1222.2298069860888	820.0
	RJ	Tecnológico	Curso	а	Distância	Gestão Hospitalar	293.8268250235524	279.0
	RJ	Licenciatura	Curso	а	Distância	História	283.4026037201119	289.0
	RJ	Licenciatura	Curso	а	Distância	Sociologia	269.552930247138	149.0
	RJ	Bacharelado	Curso	а	Distância	Ciências Contábeis	330.7527164427505	402.65
	RJ	Licenciatura	Curso	а	Distância	Pedagogia	297.7451310451306	402.65
	RJ	Bacharelado			Noturno	Administração	979.6470106030273	632.28
	RJ	Bacharelado			Noturno	Direito	1222.2298069860888	815.15
	RJ	Tecnológico	Curso	а	Distância	Gestão de Recurso	304.2915754562678	296.0
	RJ	Bacharelado	Curso	а	Distância	Administração	333.34754851326505	259.0
	RJ	Tecnológico	Curso	а	Distância	Análise e Desenvo	301.83506572768994	239.0
	RJ	Bacharelado	Curso	а	Distância	Arquitetura e Urb	674.864699462568	579.0
	RJ	Licenciatura	Curso	а	Distância	Artes VIsuais	275.44266879296515	209.0
	RJ	Licenciatura	Curso	а	Distância	Ciências Biológicas	297.1462073482199	249.0

// Aprendizado de máquina com os dados - segundo exemplo

```
//Instanciando o modelo GBTRegressor
val modelo = new GBTRegressor().setLabelCol("nota integral ampla").
setFeaturesCol("features").setMaxBins(338).setMaxIter(10)
//Removendo dados nulos
val dfFiltrada = indexed8.filter($"nota integral ampla".isNotNull)
//Treinando o modelo com os dados de treinamento
val modeloTreinado2 = modelo.fit(dfFiltrada)
// Split do dataframe indexado
val splitDF = dfFiltrada.randomSplit(Array(1,1))
val (dataframeTreinamento, dataframeTeste) = (splitDF(0), splitDF(1))
// Fazendo Predicoes com os dados de teste ou desconhecidas
val fazendoPredicoes = modeloTreinado2.transform(dataframeTeste)
// Criando avaliadores para as metricas do modelo, usando o RMSE e o R2
val avaliador = new RegressionEvaluator().setLabelCol("nota integra
l ampla").setPredictionCol("prediction").setMetricName("rmse")
val Test rmse = avaliador.evaluate(fazendoPredicoes)
val avaliador2 = new RegressionEvaluator().setLabelCol("nota integr
al ampla").setPredictionCol("prediction").setMetricName("r2")
val Test R2 = avaliador2.evaluate(fazendoPredicoes)
println("A RAIZ DO ERRO MÉDIO QUADRATICO (RMSE) nos dados de teste
= " + Test rmse)
println("O R2 nos dados de teste : " + Test R2);
// Avaliando os resultados
fazendoPredicoes.createOrReplaceTempView("minhaTabela")
spark.sql("SELECT uf busca, grau, turno, nome, prediction, nota int
egral ampla FROM minhaTabela WHERE uf busca = 'RJ'").show
```

uf_k	ousca	grau			turno	nome		prediction	nota_integral_ampl
	RJ	Bacharelado	Curso	a	Distância	Ciências Contábeis	594.	8813062569851	606.1
	RJ	Licenciatura	Curso	а	Distância	Pedagogia	574.	1427519039341	545.0
	RJ	Bacharelado	Curso	а	Distância	Enfermagem	613.	7621665644414	590.6
	RJ	Bacharelado	Curso	а	Distância	Serviço Social	570.	5937317708701	543.8
	RJ	Bacharelado			Noturno				
	RJ	Bacharelado			Noturno	Direito	653.	5714345322231	625.1
	RJ	Tecnológico	Curso	а	Distância	Gestão Hospitalar	579.	4382011417333	534.3
	RJ	Licenciatura	Curso	а	Distância	Letras - Português	577.	0800113389572	590.5
	RJ	Bacharelado	Curso	а	Distância	Ciências Contábeis	594.	8813062569851	578.9
	RJ	Bacharelado			Noturno	Administração	614.	6686201328874	595.
	RJ	Bacharelado			Noturno	Direito	653.	5714345322231	674.0
	RJ	Bacharelado			Noturno	Enfermagem	634.	1907770992116	631.2
	RJ	Tecnológico	Curso	а	Distância	Gestão de Recurso	566.	0697522518115	574.
	RJ	Bacharelado	Curso	а	Distância	Administração	594.	2880578679697	611.
	RJ	Licenciatura	Curso	а	Distância	Artes VIsuais	579.	3227745908341	609.
	RJ	Tecnológico	Curso	а	Distância	Embelezamento e I	574.	9896853953522	568.5
	RJ	Tecnológico	Curso	а	Distância	Gestão Pública	574.	9896853953522	561.0
	RJ	Tecnológico	Curso	а	Distância	Gestão de Recurso	566.	0697522518115	600.1
	RJ	Licenciatura	Curso	а	Distância	História	599.	3967928230439	619.7
	RJ	Licenciatura	Curso	а	Distância	Letras - Português	577.	0800113389572	550.