Exercício

Importaremos os recursos necessários

```
from pyspark.ml.clustering import KMeans
from pyspark.ml.feature import VectorAssembler
from pyspark.sql import SparkSession
```

Criaremos a sessão Spark.

```
spark = SparkSession.builder.appName(“app_model”).getOrCreate()
```

Carregaremos o ficheiro stocks_2021.csv num dataframe.

Com o delimitador "," e com cabeçalho

```
df = spark.read.option(“delimiter”, “,”).option(“header”, “true”).csv(“stocks_2021.csv”)
```

Modificaremos o tipo de dados das colunas open, low e close para float.

```
df = df.withColumn(‘open’, df.open.cast(‘float’))
df = df.withColumn(‘low’, df.low.cast(‘float’))
df = df.withColumn(‘close’, df.close.cast(‘float’))
```

Criaremos uma coluna 'features', utilizando as colunas open, low e close através do VectorAssembler.

```
va = VectorAssembler(inputCols=[‘open’,’low’,’close’], outputCol=’features’)
```

Aplicamos o VectorAssembler ao DataFrame

```
va_df = va.transform(df)
```

Criaremos o objeto K-means e configurá-lo-emos para estabelecer 5 clusters.

```
kmeans = KMeans(k=5)
```

Treinaremos o modelo com base em “features”

```
model = kmeans.fit(va_df.select(‘features’))
```

Aplicaremos o modelo KMeans ao DataFrame

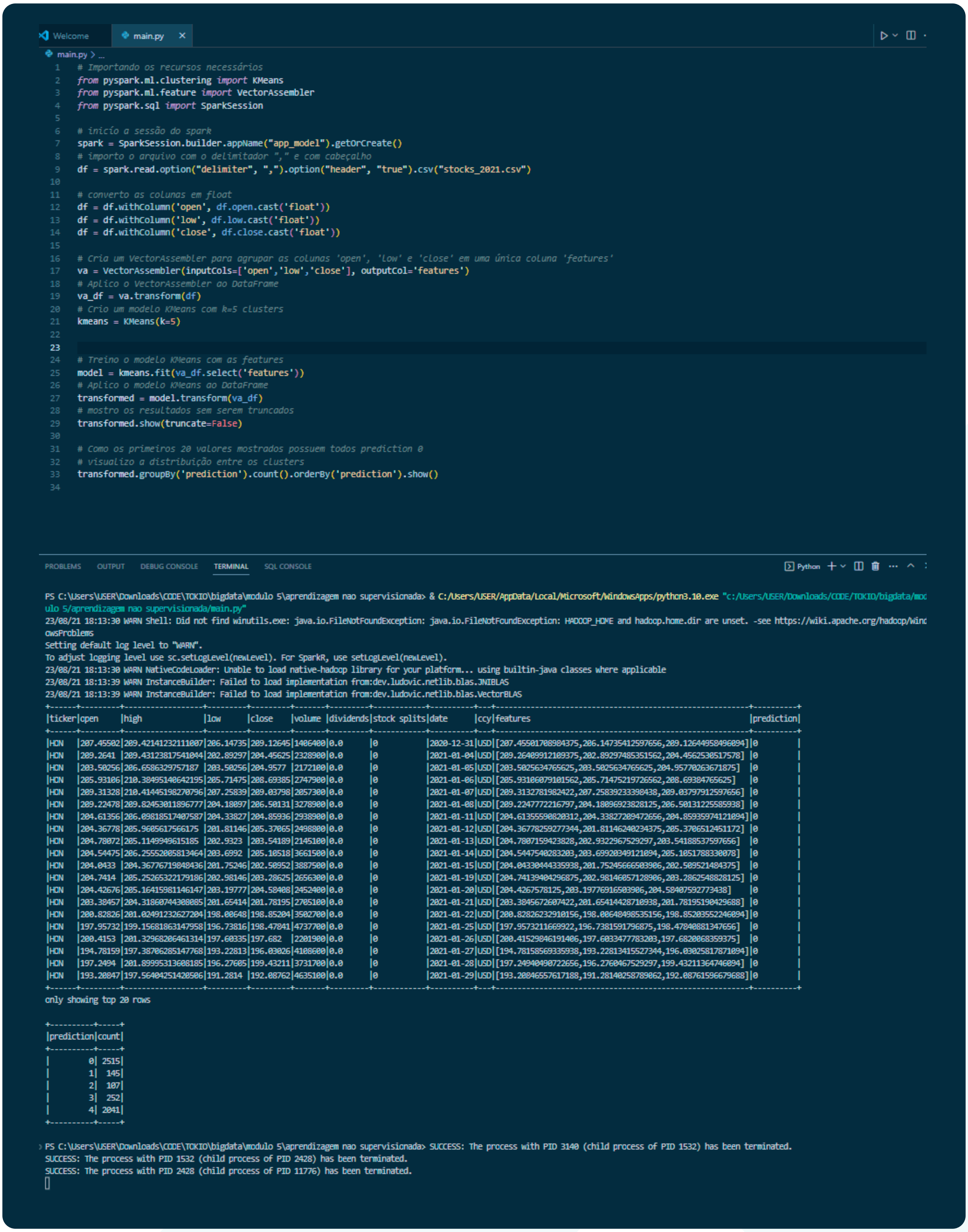
```
transformed = model.transform(va_df)
```

Mostramos os resultados sem serem truncados

```
transformed.show(truncate=False)
```

Como os primeiros 20 valores mostrados possuem todos prediction 0,
Podemos visualizar a distribuição entre os clusters para ter uma melhor idéia.

```
transformed.groupBy(‘prediction’).count().orderBy(‘prediction’).show()
```



```python
# Importando os recursos necessários
from pyspark.ml.clustering import KMeans
from pyspark.ml.feature import VectorAssembler
from pyspark.sql import SparkSession

# inicío a sessão do spark
spark = SparkSession.builder.appName("app_model").getOrCreate()
# importo o arquivo com o delimitador "," e com cabeçalho
df = spark.read.option("delimiter", ",").option("header", "true").csv("stocks_2021.csv")

# converto as colunas em float
df = df.withColumn('open', df.open.cast('float'))
df = df.withColumn('low', df.low.cast('float'))
df = df.withColumn('close', df.close.cast('float'))

# Cria um VectorAssembler para agrupar as colunas 'open', 'low' e 'close' em uma única coluna 'features'
va = VectorAssembler(inputCols=['open','low','close'], outputCol='features')
# Aplico o VectorAssembler ao DataFrame
va_df = va.transform(df)
# Crio um modelo KMeans com k=5 clusters
kmeans = KMeans(k=5)


# Treino o modelo KMeans com as features
model = kmeans.fit(va_df.select('features'))
# Aplico o modelo KMeans ao DataFrame
transformed = model.transform(va_df)
# mostro os resultados sem serem truncados
transformed.show(truncate=False)

# Como os primeiros 20 valores mostrados possuem todos prediction 0
# visualizo a distribuição entre os clusters
transformed.groupBy('prediction').count().orderBy('prediction').show()

```

```
PS C:\Users\USER\Downloads\CODE\TOKIO\bigdata\modulo 5\aprendizagem nao supervisionada> & C:/Users/USER/AppData/Local/Microsoft/WindowsApps/python3.10.exe "c:/Users/USER/Downloads/CODE/TOKIO/bigdata/mod
ulo 5/aprendizagem nao supervisionada/main.py"
23/08/21 18:13:30 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: java.io.FileNotFoundException: HADOOP_HOME and hadoop.home.dir are unset. -see https://wiki.apache.org/hadoop/Wind
owsProblems
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
23/08/21 18:13:30 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
23/08/21 18:13:39 WARN InstanceBuilder: Failed to load implementation from:dev.ludovic.netlib.blas.JNIBLAS
23/08/21 18:13:39 WARN InstanceBuilder: Failed to load implementation from:dev.ludovic.netlib.blas.VectorBLAS
+------+---------+------------------+---------+---------+-------+---------+------------+----------+---+----------------------------------------------------------+----------+
|ticker|open     |high              |low      |close    |volume |dividends|stock splits|date      |ccy|features                                                  |prediction|
+------+---------+------------------+---------+---------+-------+---------+------------+----------+---+----------------------------------------------------------+----------+
|HON   |207.45502|209.42141232111007|206.14735|209.12645|1406400|0.0      |0           |2020-12-31|USD|[207.45501708984375,206.14735412597656,209.12644958496094]|0         |
|HON   |209.2641 |209.43123817541044|202.89297|204.45625|2328900|0.0      |0           |2021-01-04|USD|[209.26409912109375,202.89297485351562,204.4562530517578] |0         |
|HON   |203.50256|206.6586329757187 |203.50256|204.9577 |2172100|0.0      |0           |2021-01-05|USD|[203.5025634765625,203.5025634765625,204.95770263671875]  |0         |
|HON   |205.93106|210.38495140642195|205.71475|208.69385|2747900|0.0      |0           |2021-01-06|USD|[205.93106079101562,205.71475219726562,208.69384765625]   |0         |
|HON   |209.31328|210.41445198270796|207.25839|209.03798|2057300|0.0      |0           |2021-01-07|USD|[209.3132781982422,207.25839233398438,209.03797912597656] |0         |
|HON   |209.22478|209.82453011896777|204.18097|206.50131|3278900|0.0      |0           |2021-01-08|USD|[209.2247772216797,204.18096923828125,206.50131225585938] |0         |
|HON   |204.61356|206.09818517407587|204.33827|204.85936|2938900|0.0      |0           |2021-01-11|USD|[204.61355590820312,204.33827209472656,204.85935974121094]|0         |
|HON   |204.36778|205.9605617566175 |201.81146|205.37065|2498800|0.0      |0           |2021-01-12|USD|[204.36778259277344,201.81146240234375,205.3706512451172] |0         |
|HON   |204.78072|205.1149949615185 |202.9323 |203.54189|2145100|0.0      |0           |2021-01-13|USD|[204.7807159423828,202.9322967529297,203.54188537597656]  |0         |
|HON   |204.54475|206.25552005813464|203.6992 |205.10518|3661500|0.0      |0           |2021-01-14|USD|[204.5447540283203,203.69920349121094,205.1051788330078]  |0         |
|HON   |204.0433 |204.36776719848436|201.75246|202.50952|3887500|0.0      |0           |2021-01-15|USD|[204.04330444335938,201.75245666503906,202.509521484375]  |0         |
|HON   |204.7414 |205.25265322179186|202.98146|203.28625|2656300|0.0      |0           |2021-01-19|USD|[204.74139404296875,202.98146057128906,203.2862548828125] |0         |
|HON   |204.42676|205.16415981146147|203.19777|204.58408|2452400|0.0      |0           |2021-01-20|USD|[204.4267578125,203.19776916503906,204.58407592773438]    |0         |
|HON   |203.38457|204.31860744308085|201.65414|201.78195|2705100|0.0      |0           |2021-01-21|USD|[203.3845672607422,201.65414428710938,201.78195190429688] |0         |
|HON   |200.82826|201.02491232627204|198.00648|198.85204|3502700|0.0      |0           |2021-01-22|USD|[200.82826232910156,198.00648498535156,198.85203552246094]|0         |
|HON   |197.95732|199.15681863147958|196.73816|198.47841|4737700|0.0      |0           |2021-01-25|USD|[197.9573211669922,196.7381591796875,198.47840881347656]  |0         |
|HON   |200.4153 |201.32968206461314|197.60335|197.682  |2201900|0.0      |0           |2021-01-26|USD|[200.41529846191406,197.6033477783203,197.6820068359375]  |0         |
|HON   |194.78159|197.38706285147768|193.22813|196.03026|4108600|0.0      |0           |2021-01-27|USD|[194.78158569335938,193.22813415527344,196.03025817871094]|0         |
|HON   |197.2494 |201.89995313608185|196.27605|199.43211|3731700|0.0      |0           |2021-01-28|USD|[197.24940490722656,196.2760467529297,199.43211364746094] |0         |
|HON   |193.20847|197.56404251420506|191.2814 |192.08762|4635100|0.0      |0           |2021-01-29|USD|[193.20846557617188,191.28140258789062,192.08761596679688]|0         |
+------+---------+------------------+---------+---------+-------+---------+------------+----------+---+----------------------------------------------------------+----------+
only showing top 20 rows

+----------+-----+
|prediction|count|
+----------+-----+
|         0| 2515|
|         1|  145|
|         2|  107|
|         3|  252|
|         4| 2041|
+----------+-----+

PS C:\Users\USER\Downloads\CODE\TOKIO\bigdata\modulo 5\aprendizagem nao supervisionada> SUCCESS: The process with PID 3140 (child process of PID 1532) has been terminated.
SUCCESS: The process with PID 1532 (child process of PID 2428) has been terminated.
SUCCESS: The process with PID 2428 (child process of PID 11776) has been terminated.
▯
```