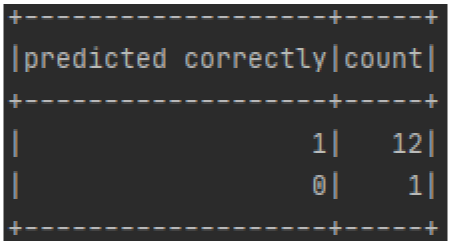
Big Data Final Project

5/4/2020

Christopher Fulton

For the final project, I utilized pyspark's machine learning functions and pyspark's SQL data types and dataframes in order to predict, given parameters from the FBI crime statistics data, whether a particular year will have a higher or lower than average murder and manslaughter rate.

- The program uses SQL dataframes to store the raw data found within a standard CSV file containing FBI statistics. A function is run to determine whether each year has a high murder and manslaughter rate and stores that within the SQL dataframe.
- A modified dataframe is created using all data except the given murder rate in order for the machine learning functions to properly train and predict.
- Vectors are created and transformed to split the data usage and prepare it for the ML functions; One set is used to train the program into making accurate predictions, the other is the test the accuracy after training.
- Using pyspark.ml, a regression analysis is performed, where the program analyzes the given data and assesses the conditions required for a statistically likely high murder rate. (Similar to the program SCALA used in econometrics)
- The program then tests itself with the remaining data set and outputs it's correct and incorrect prediction values:



```
+-------------------+-----+
|predicted correctly|count|
+-------------------+-----+
|                  1|   12|
|                  0|    1|
+-------------------+-----+
```