This Assignment is aimed at comparing the performance of several algorithm models for different programming languages such as Scala and Python on Spark ML.

- ***PC Features:***

  -Windows 10 Pro 64 bits

  -Processor: Intel(R) Core(TM) i5-4460 CPU @ 3.2 GHz 3.2 GHz

  -16 Gb RAM

  -Hard Drive SSD 500 Gb

  -Java JDK 1.8.0_161

  -Spark 2.3.0

  -IDE Jetbrains Pycharm 2018.1

  -IDE Intellij 2018.1

  -Scala version 2.11.12

For this, we have created a Dataset to train the different algorithms models with a generator of data program in python for clustering and Classification test time and performance.

Results of the test:

- Computing Time Based on Bisecting-K-means Algorithm.

Nº Cores	Clustering Algorithm		
	Bisecting-K-means		
	Scala Time	Python Time	Difference Time (%)
1	194	209	7%
2	108	121	11%
3	81	95	15%
4	68	87	22%

Average Performance Scala-Python 14%

- Computing Time Based on K-means Algorithm.

Nº Cores	Clustering Algorithm		
	Kmeans		
	Scala Time	Python Time	Difference Time (%)
1	53	70	24%
2	39	43	9%
3	30	34	12%
4	26	28	7%

Average Performance Scala-Python 13%

- Computing Time Based on LogisticRegression Algorithm.

Nº Cores	Classification Algorithm		
	LogisticRegression		
	Scala Time	Python Time	Difference Time (%)
1	62	76	18%
2	39	41	5%
3	30	30	0%
4	23	30	23%

Average Performance Scala-Python 12%

- Computing Time Based on DecisionTree Algorithm.

Nº Cores	Classification Algorithm		
	DecisionTree		
	Scala Time	Python Time	Difference Time (%)
1	214	270	21%
2	126	161	22%
3	103	126	18%
4	81	115	30%

Average Performance Scala-Python 23%

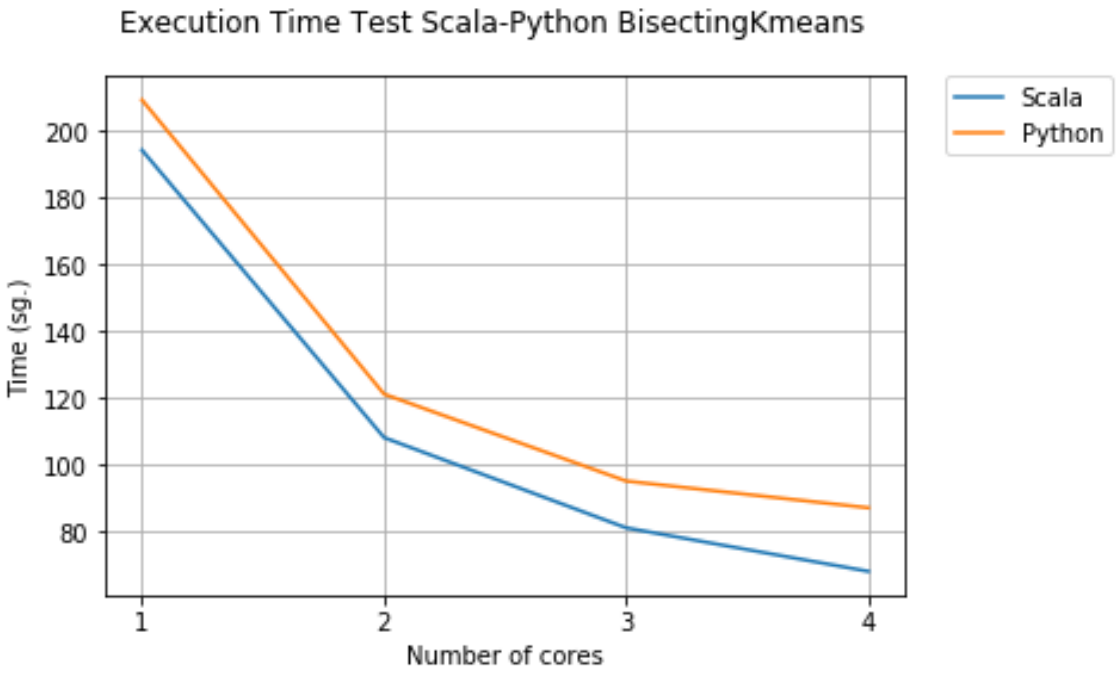
- Computing Time Based on RandomForest Algorithm.

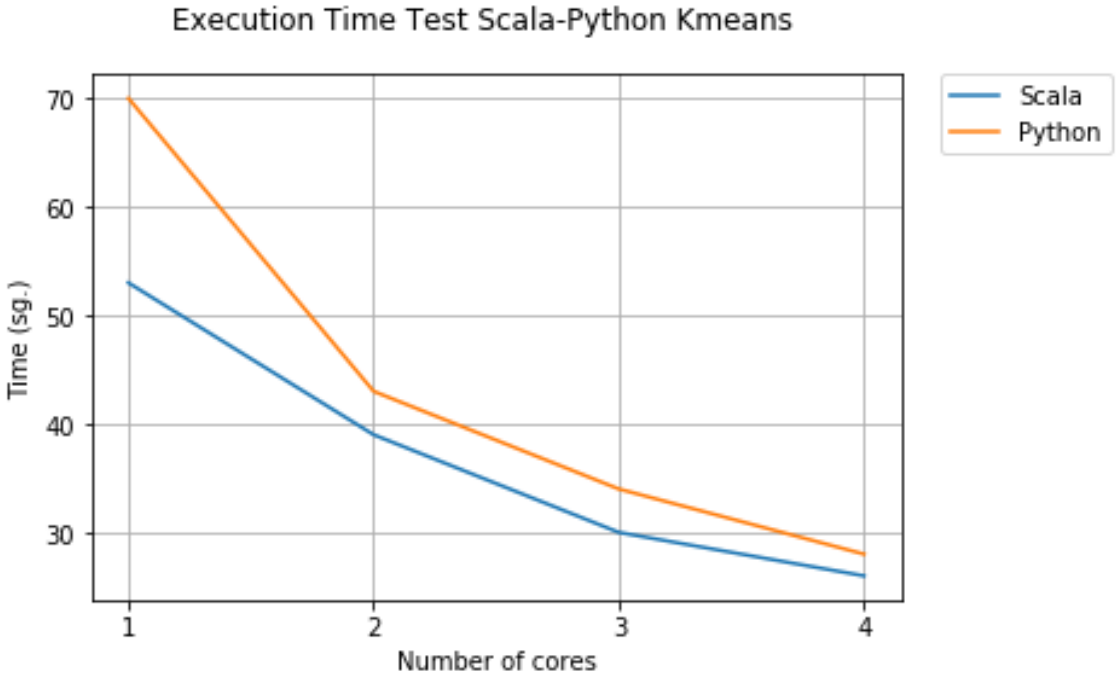
Nº Cores	Classification Algorithm		
	RandomForest		
	Scala Time	Python Time	Difference Time (%)
1	207	283	27%
2	139	158	12%
3	106	110	4%
4	81	104	22%

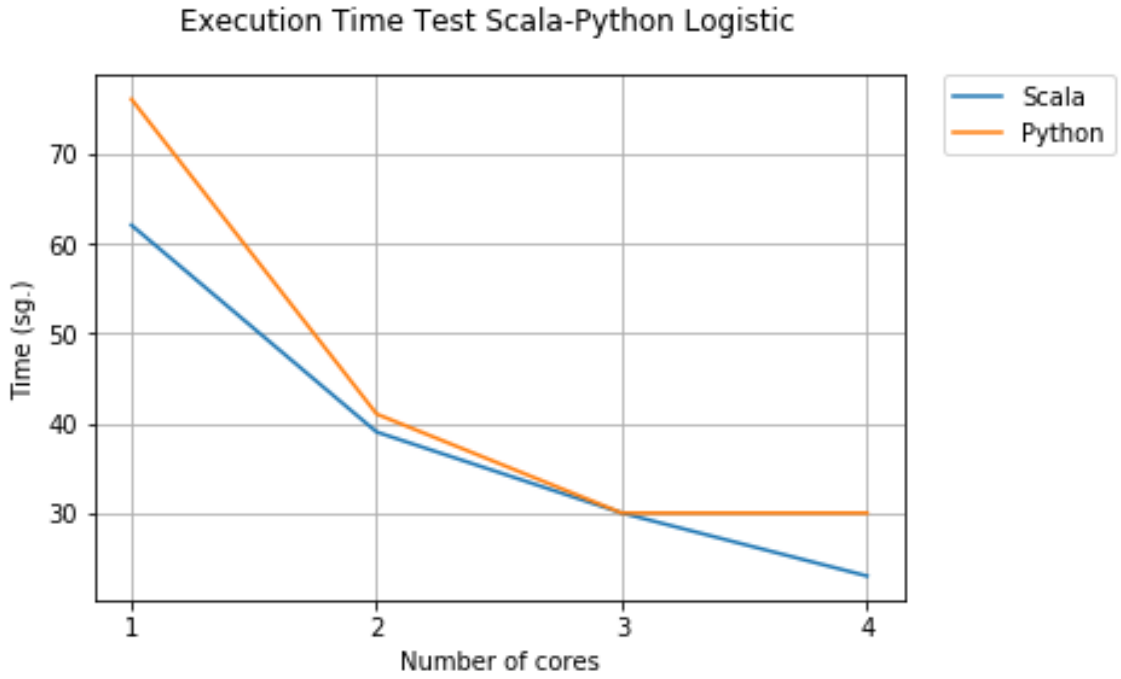
Average Performance Scala-Python 16%

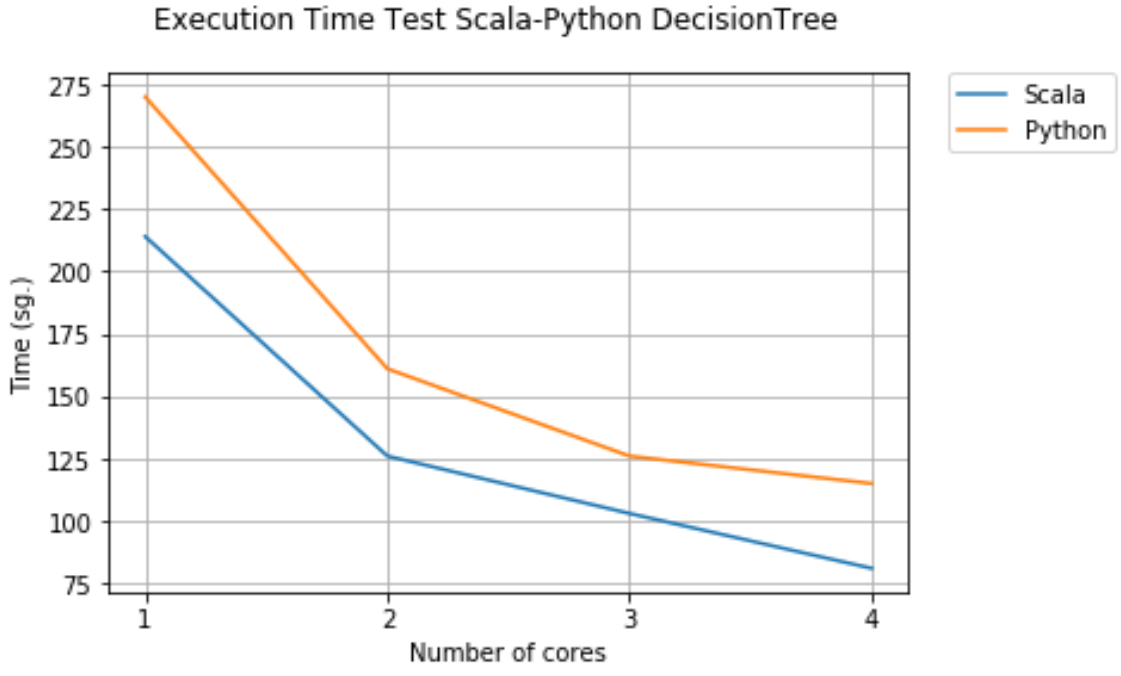
Global Average Performance Scala-Python 15%

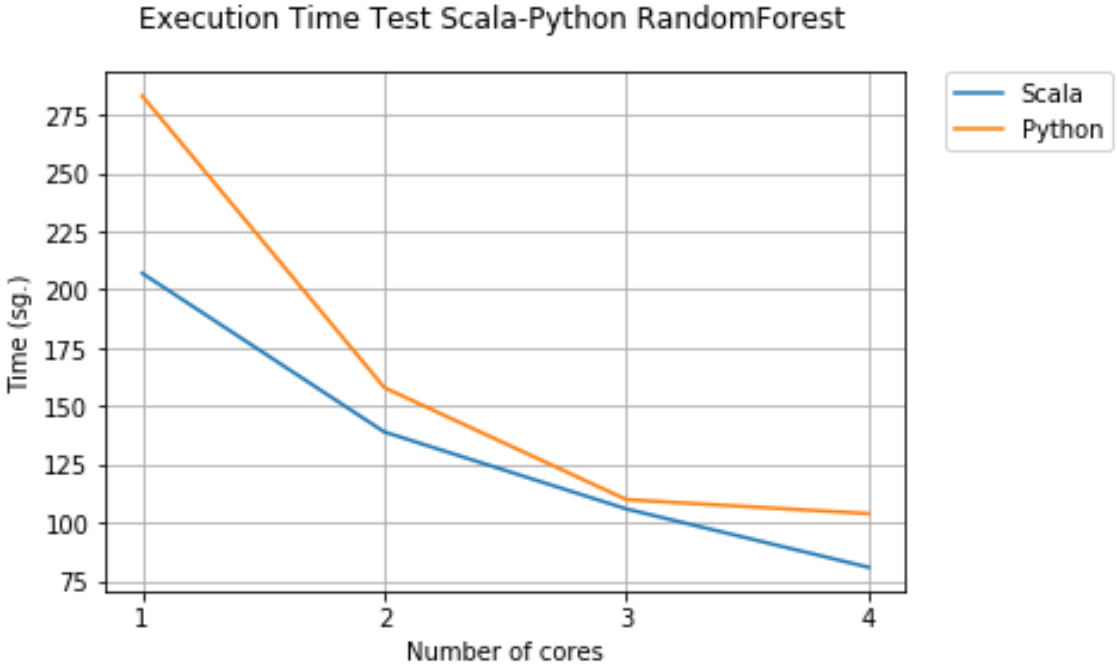
- Computing Time Based on Graphs.





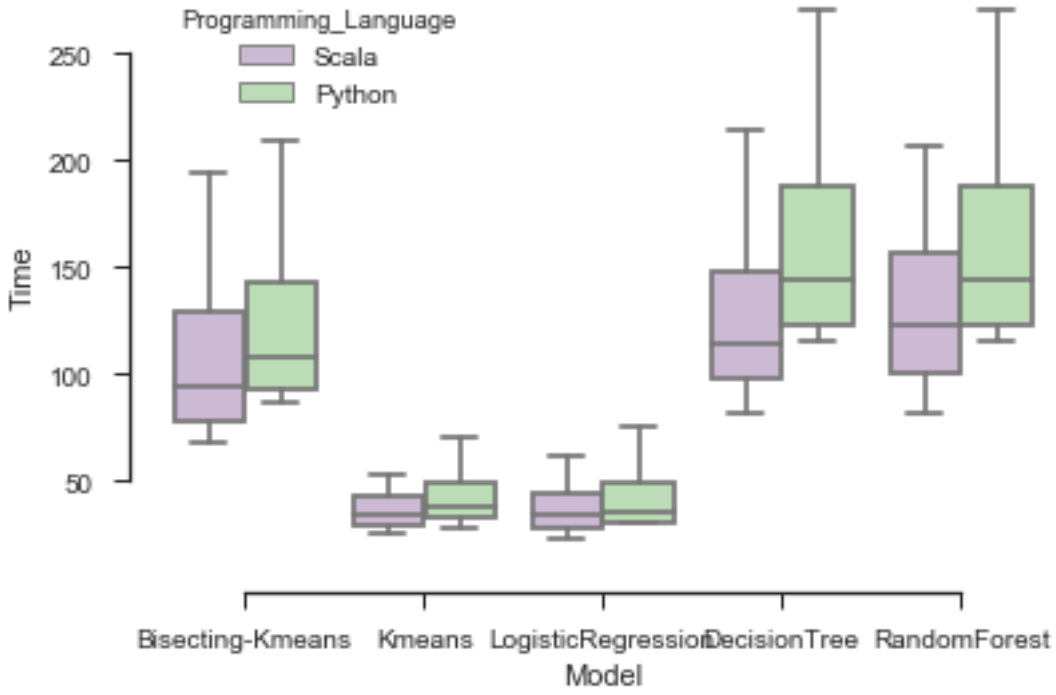

Execution Time Test Scala-Python Logistic
Scala
Python
Time (sg.)
70
60
50
40
30
1
2
3
4
Number of cores


Execution Time Test Scala-Python DecisionTree
Scala
Python
Time (sg.)
275
250
225
200
175
150
125
100
75
1
2
3
4
Number of cores

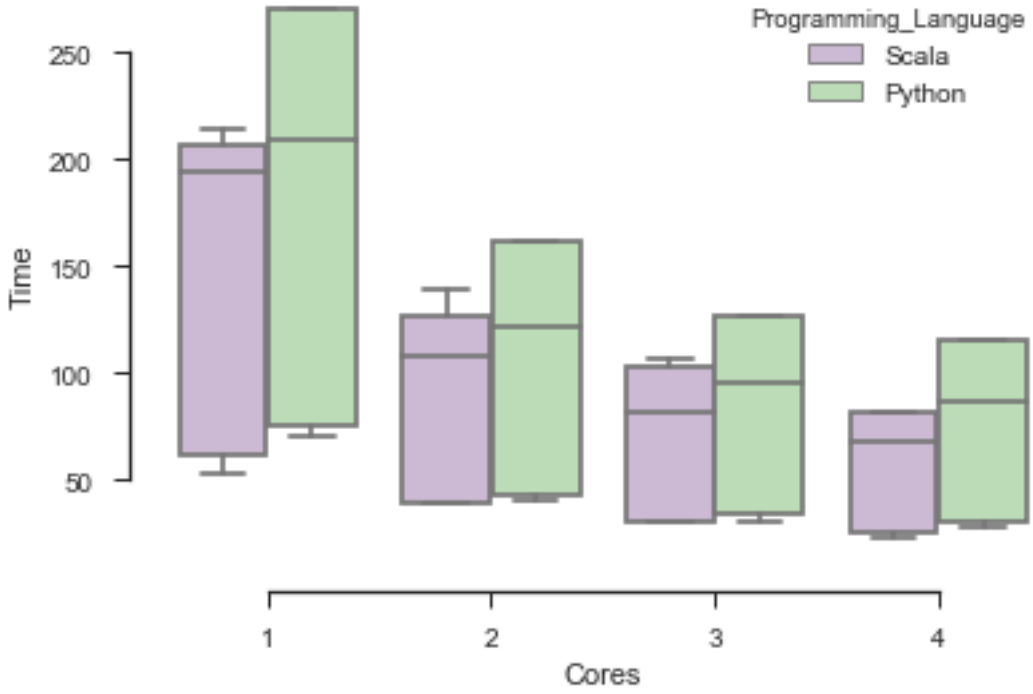

Execution Time Test Scala-Python RandomForest
Time (sg.)
275
250
225
200
175
150
125
100
75
1
2
3
4
Number of cores
Scala
Python

Analysis of the test:

- Analysis of the computing Time on Boxplot Graph by Algorithms Models.



- Analysis of the computing Time on Boxplot Graph by Cores (it take account of all models.)



Test of Wilcoxon:

*The **Wilcoxon signed-rank test** is a non-parametric statistical hypothesis test used to compare two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test). It can be used as an alternative to the paired Student's t-test, t-test for matched pairs, or the t-test for dependent samples when the population cannot be assumed to be normally distributed. A Wilcoxon signed-rank test is a nonparametric test that can be used to determine whether two dependent samples were selected from populations having the same distribution.*

Reference: http://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test

Hypothesis:

H_0: ***Median*** *Scala-Times ≥* ***Median*** *Python-Times.*

H_1: ***Median*** *Scala-Times <* ***Median*** *Python-Times.*

Value of p:

P = 0.00013 < 0.05 *Therefore, the Hypothesis null is refused.*

Conclusion:

We can say that the performance with Programming_Language Scala is better than Programming_Language Python.

NOTE: You can see the code in WilcoxonTest.py

Final Conclusion:

Afterwards, we can do some observations:

1. *In general, Scala has about a 15% better global Average performance than Python.*

2. *Classification DecisionTree Algorithm model has obtained the best performance with a 23% less of time.*

3. *Classification LogisticRegression Algorithm model has obtained the worst performance with a 12% less of time.*

4. *It's better to use programming_Language Scala to train whatever model and especially for DecisionTree and RandomForest. (you look at boxplot Analysis of the computing Time on Boxplot Graph by Algorithms Models).*

5. *The performance of Programming_Language Scala (15% time-saving) regard to Programming_Language Python is independent of the hardware.*