Nile University

School of Communication and Information Technology (CIT)

Master of Science/Engineering MSCIT/ MCIT Program

Status Report 20

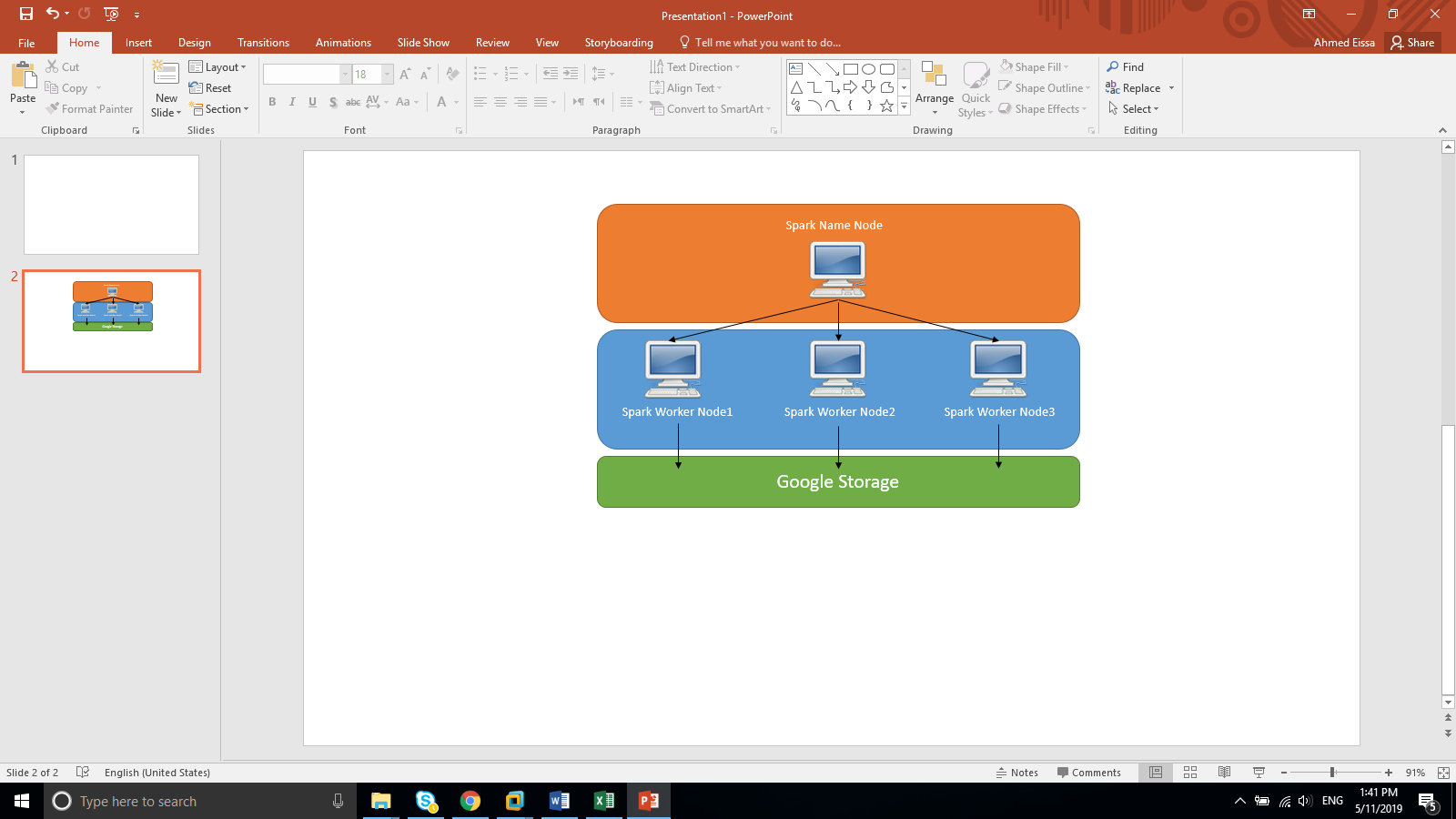
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version No.** | **Date** | **Description** | **Created By** | **e-mail** |
| 0.1 | 11-05-2019 | Status Report 29 | Ahmed Mohamed Abdel Rahman | [Robot209@gmail.com](mailto:Robot209@gmail.com) |
|  |  |  |  |  |

## Objective

* Build a Knowledgebase contains a group of datasets characteristics & classification accuracy for different classifiers
* Using this knowledgebase to select the best classifier for any given dataset (based on its characteristics)

## Environemnt

* I have used Spark Cluster on Google Cloud
* The Cluster consists of:
  + One name node (2 CPU, 13 GB RAM, 500GB HDD)
  + Three Worker nodes (2 CPU, 13 GB RAM, 500GB HDD)
* I have used Spark 2.3 and Scala 2.11



## Datasets

* I have used 73 different datasets from different sources
* The datasets have (large, medium and small) number of features, (large, medium and small) number of instances, datasets with missing values or with no missing values

## extracted Characteristics

* I have extracted the below characteristic for each dataset:
  + **For the Dataset**
    - Number of instances
    - Log (Number of instances)
    - Number of Features
    - Log (Number of Features)
    - Number of Classes
    - Number of numerical Features
    - Number of Categorical Features
    - Ratio between Categorical & Numerical
    - Number of Instances to Number of Features ratio
    - Number of Missing Value
    - Ratio of missing value
  + **For Classes**
    - Class Entropy
    - Classes Probabilities:
      * Minimum
      * Maximum
      * Standard deviation
      * Mean
  + **For Categorical Features**
    - Sum of all symbols
    - Mean of all symbols
    - Standard Deviation for all symbols
  + **For Numerical Features**
    - Skewness for all Numerical features
      * Max
      * Min
      * Standard Deviation
      * Mean
    - Kurtosis for all Numerical features
      * Max
      * Min
      * Standard Deviation
      * Mean

## Data Processing

* All the datasets have been processed as following:
  + **String Encoding for Categorical Features**: encodes a string column of labels to a column of label indices. The indices are in [0, numLabels)

Ref: <https://spark.apache.org/docs/latest/ml-features.html#stringindexer>

* + **Standard Scaling**: using “Standard Scalar” to normalizing each feature to have unit standard deviation and/or zero mean

Ref: <https://spark.apache.org/docs/latest/ml-features.html#standardscaler>

* + **Min-Max Scaling:** rescaling each feature to a specific range (often [0, 1])

ref: <https://spark.apache.org/docs/latest/ml-features.html#minmaxscaler>

* + **Replacing missing value:** The Imputer estimator completes missing values in a dataset, either using the mean or the median of the columns in which the missing values are located

Ref: <https://spark.apache.org/docs/latest/ml-features.html#imputer>

* + **Vector Assembler:** combines a given list of columns into a single vector column

Ref: <https://spark.apache.org/docs/latest/ml-features.html#vectorassembler>

## Used Classifiers

* I have used 9 classifiers, 7 classifiers from Spark MLlib and developed 2 additional custom classifiers, so the used classifiers are:
  + Random Forest
  + Logistic Regression
  + Decision Tree
  + Multilayer Perceptron
  + Linear SVC
  + Naïve Bayes
  + GBT
  + LDA (Custom)
  + QDA (Custom)
* Some classifier only supports binary classes [Linear SVC & GBT] and some support binary and multiple classes
* I have used Spark MLlib “MulticlassClassificationEvaluator” class with “Accuracy” metric to select the best classifier
* I have used default hyperparameters values for all Algorithms (using Spark MLlib default values)

## KB Classification

* The datasets characteristic will be considered as the features and best algorithm will be considered as the label
* I have tried the same 9 classifiers to select which classifier able to select the best algorithm based on the Dataset characteristics.
* Result:
  + RandomForestClassifier:0.5
  + LogisticRegression:0.1875
  + DecisionTreeClassifier:0.5
  + MultilayerPerceptronClassifier:0.3125
  + LinearSVC:0.0
  + NaiveBayes:0.36363636363636365
  + GBTClassifier:0.0
  + LDA:0.3125
  + QDA:0.0625

## Appendix

* **Excel File “Status report 20” contains three sheets:**
  + Datasets Sources
  + Knowledge base
  + Time Log