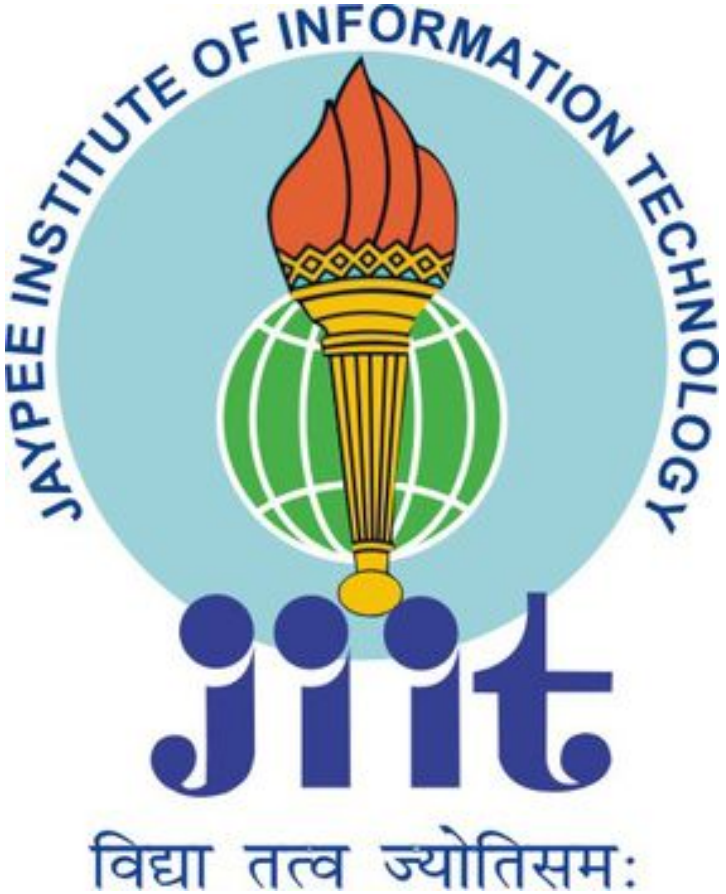
JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY

ALGORITHMS PROJECT



By-
Aditya Lahiri - 15103138
Nipun Mittal - 15103303
Yash Singhal - 15103315
Batch - B7

LITERACY CALCULATOR - A STUDY OF INDIAN DEMOGRAPHY

OBJECTIVE:

To establish a predictive model(s) to decipher a relationship between various local features and the corresponding literate population in a given district in India.

THEORY:

An up and coming field in Computer Science is data analytics and manipulation. Hence, we decided to undertake a project in the field for our algorithms project.
It highlights the basic concepts of **Machine Learning** including **Supervised learning** and **Prediction techniques.** Moreover, we decided to learn the **Scala programming language(a functional programming language)** and the **Spark engine** in hopes of future use. Most common PCs are ill equipped to handle the spark engine and hence the spark part of the project is implemented in hypothesis only.

The following libraries in python were used-

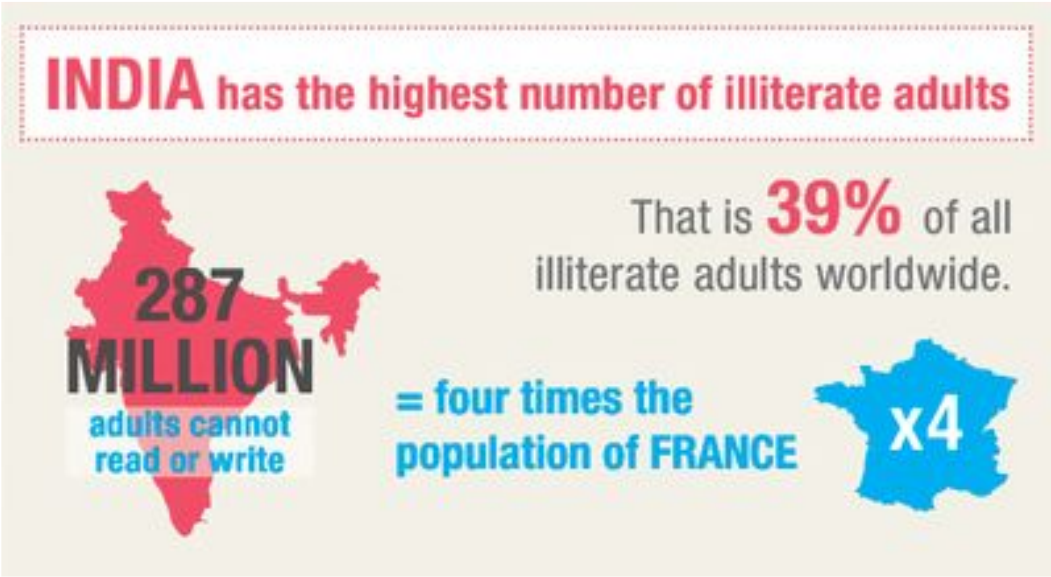
- Pandas
- CSV
- Numpy
- Matplotlib
- sklearn

The following imports were used in Spark-

```
import org.apache.spark.ml.{Pipeline, PipelineModel}
```

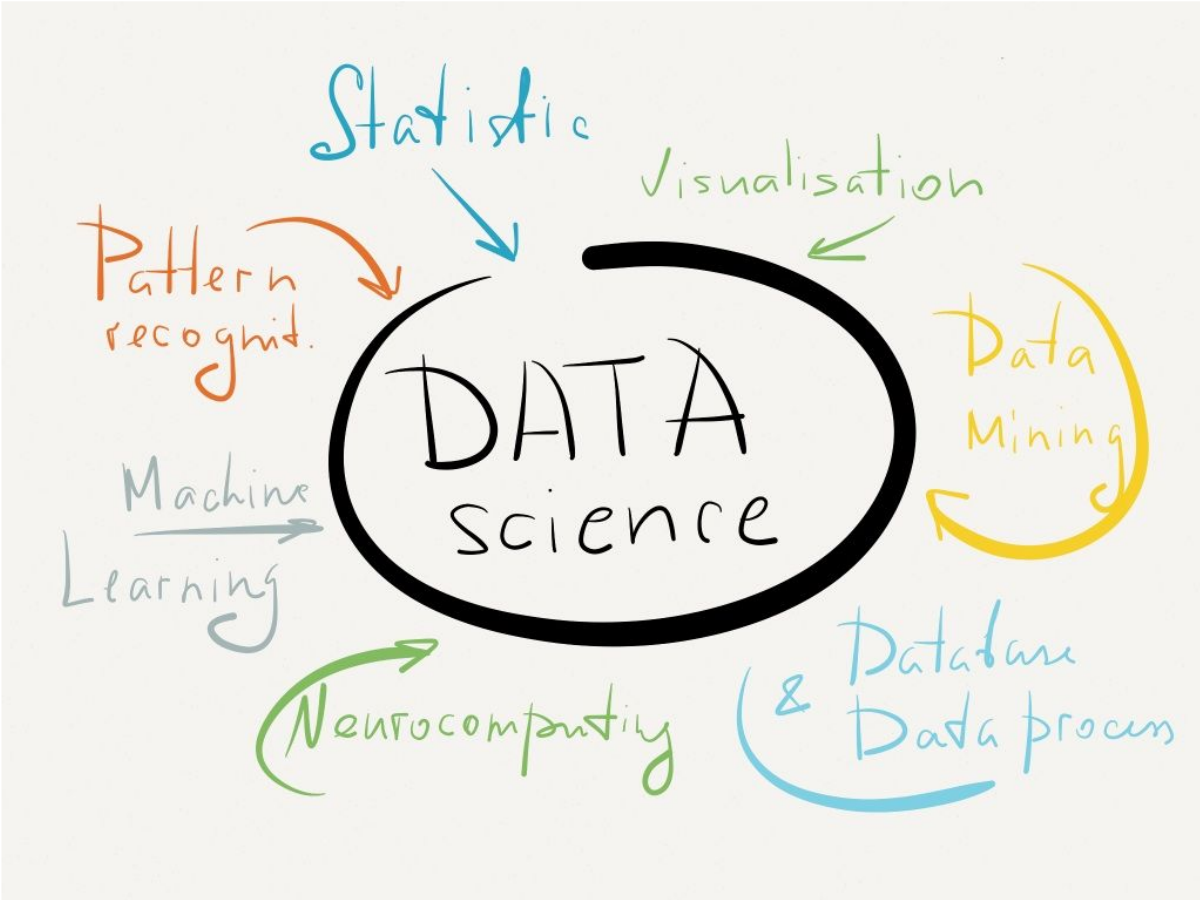
```
import org.apache.spark.ml.feature.{HashingTF, Tokenizer,
VectorAssembler, VectorIndexer, StringIndexer}
import org.apache.spark._
import org.apache.spark.SparkContext._
import org.apache.log4j._
import org.apache.spark.sql.functions._
import org.apache.spark.ml.classification.NaiveBayes
import
org.apache.spark.ml.evaluation.MulticlassClassificationEvaluator
import org.apache.spark.ml.classification.LogisticRegression
```

The problem of Indian literacy-



India is a country with the highest rate of growth of GDP in the world as well as the overall third highest GDP in the world. Hence, the future of India as a super

power is all but secured. However, one vital obstacle lies in the path of our great country: **LITERACY.** The literacy rate of India is abysmal when compared to other developed and developing countries which universally have a literacy rate of over 90%, while we are at a mere 75 -80 %. The aim of the project is to assess this problem and to suggest a strategy to tackle the problem. However, a comprehensive study is impossible due to given constraints of time and resources. Hence, our aum is to scratch the surface and ignite the flame that is a data analysis approach to eradicading poverty in India.



OVERVIEW:

In order to carry out the following project we went through the following steps-

- **Crawled the relevant data**
  - Tools used :
    - R Programming language
    - Python (pdf, csv and xl libraries)
  - Downloaded pdf files district wise from censusindia.gov.in using Python.
  - Parse all pdf files into xlsx format
  - Then combine all district wise xlsx files into one using Python
  - Then convert all xlsx files into csv files using R and combine them to one named "census.csv" using Python.
  - A total of close to **800** records were obtained (individual districts)
  - Appropriate operations were carried out to remove incomplete and non-numeric records using python pandas data frame operations.
  - The final tally of records was at **~470**.
- **Analysis of data**
  - A overview of the data was carried out. The total variables were:

Unnamed: 0	249 non-null int64
Unnamed: 0.1	249 non-null int64
State	249 non-null object
District	249 non-null object
Persons	249 non-null float64

Males 249 non-null float64
Females 249 non-null float64
Growth (1991 - 2001) 249 non-null object
Rural 249 non-null object
Urban 249 non-null object
Scheduled Caste population 249 non-null object
Percentage - SC to total 249 non-null object
Number of households 249 non-null float64
Household size (per household) 249 non-null float64
Sex ratio (females per 1000 males) 249 non-null float64
Sex ratio (0-6 years) 249 non-null float64
Scheduled Tribe population 249 non-null object
Percentage to total population (ST) 249 non-null object
Persons- literate 249 non-null float64
Males- Literate 249 non-null float64
Females- Literate 249 non-null float64
Persons- literacy rate 249 non-null float64
Males- Literatacy Rate 249 non-null float64
Females- Literacy Rate 249 non-null float64
Total Educated 249 non-null float64
Data without level 249 non-null float64
Below Primary 249 non-null float64
Primary 249 non-null float64
Middle 249 non-null float64
Matric/Higher Secondary/Diploma 249 non-null float64

Graduate and Above 249 non-null float64
0 - 4 years 249 non-null float64
5 - 14 years 249 non-null float64
15 - 59 years 249 non-null float64
60 years and above (Incl. A.N.S.) 249 non-null float64
Total workers 249 non-null float64
Main workers 249 non-null float64
Marginal workers 249 non-null float64
Non-workers 249 non-null float64
SC#1 Name 249 non-null object
SC#1 Population 249 non-null float64
SC#2 Name 249 non-null object
SC#2 Population 249 non-null float64
SC#3 Name 249 non-null object
SC#3 Population 249 non-null float64
Religeon#1 Name 249 non-null object
Religeon#1 Population 249 non-null float64
Religeon#2 Name 249 non-null object
Religeon#2 Population 249 non-null float64
Religeon#3 Name 249 non-null object
Religeon#3 Population 249 non-null float64
ST#1 Name 249 non-null object
ST#1 Population 249 non-null object
ST#2 Name 249 non-null object
ST#2 Population 249 non-null object
ST#3 Name 249 non-null object
ST#3 Population 249 non-null object

Imp Town#1 Name 249 non-null object
Imp Town#1 Population 249 non-null float64
Imp Town#2 Name 249 non-null object
Imp Town#2 Population 249 non-null float64
Imp Town#3 Name 249 non-null object
Imp Town#3 Population 249 non-null float64
Total Inhabited Villages 249 non-null float64
Drinking water facilities 249 non-null float64
Safe Drinking water 249 non-null float64
Electricity (Power Supply) 249 non-null float64
Electricity (domestic) 249 non-null object
Electricity (Agriculture) 249 non-null object
Primary school 249 non-null float64
Middle schools 249 non-null object
Secondary/Sr Secondary schools 249 non-null object
College 249 non-null object
Medical facility 249 non-null object
Primary Health Centre 249 non-null object
Primary Health Sub-Centre 249 non-null object
Post, telegraph and telephone facility 249 non-null float64
Bus services 249 non-null float64

Paved approach road	249 non-null float64
Mud approach road	249 non-null object
Permanent House	249 non-null float64
Semi-permanent House	249 non-null float64
Temporary House	249 non-null float64

- Operations were carried out to calculate the Pearson Correlation coefficient of Persons- literate and the other variable using a Correlation matrix. We narrowed down on a list of highly correlated columns:

Persons	469 non-null float64
Males	469 non-null float64
Females	469 non-null float64
Number of households	469 non-null float64
Males- Literate	469 non-null float64
Females- Literate	469 non-null float64
Total Educated	469 non-null float64
Below Primary	469 non-null float64
Primary	469 non-null float64
Middle	469 non-null float64
Matric/Higher Secondary/Diploma	469 non-null float64
Graduate and Above	469 non-null float64
0 - 4 years	469 non-null float64
5 - 14 years	469 non-null float64
15 - 59 years	469 non-null float64
60 years and above (Incl. A.N.S.)	469 non-null float64
Total workers	469 non-null float64
Main workers	469 non-null float64
Non-workers	469 non-null float64
Religeon#1 Population	469 non-null float64

Imp Town#2 Population	469 non-null float64
Imp Town#3 Population	469 non-null float64

- Now, we obtained a new data frame containing only the above mentioned columns.
- We divided the data frame into two- test and train
- We further divided the data frame into - xtrain, ytrain, xtest, ytest
- We based our predictive model on the following ML algorithms and obtained the corresponding percentage accuracy in prediction -
  - ➢ Random Decision Forests - 0.87104937564043805
  - ➢ K Nearest Neighbor(KNN) - 0.91215628552481975
  - ➢ Gaussian Naive Bayes - 0.91215628552481975

MISCELLANEOUS

The future of big data lies in the use of clusters. In order to honor this trend, we decided to undertake the same operations using the **SPARK engine.** However, our given resources proved to be insufficient for the given operations and hence we were unable to verify our results. However, we are confident that with superior hardware, we can achieve our desired results.

CONCLUSION:

It is possible to predict the literacy rate of an area on the basis of various characteristics of the local demography. Hence, this could be a vital first step in the application of data analytics and machine learning in order to solve the problem of illiteracy in India. Investments should be made according to the calculations made from the results of this project to obtain optimum results.

IMPROVEMENTS:

The original aim of the project was not just to predict the literacy rate, but to highlight the distribution of resources required in order to obtain a 100 % literacy rate in an area. However, this required us to reverse engineer the independent variables from the dependent variable. The mathematics of this is extremely complex and beyond the purview of this project.

REFERENCES:

- www.udemy.com
- www.coursera.org
- docs.scala-lang.org
- spark.apache.org/documentation.html
- matplotlib.org/
- scikit-learn.org/stable/documentation.html
- pandas.pydata.org
- https://stackoverflow.com
- https://bigdatauniversity.com
- www.numpy.org

THE END

Thank You.