Data Analysis and Data Representation using Python

Rupika[1] Shalini[2] Kirti Bhatia[3]
[1]M.Tech Scholar [2,3]Assistant Professor
[1,2,3]Department of Computer Science and Engineering
[1,2,3]SKITM, Haryana, India

Abstract— Data analysis and data visualization is the requirement of today's organization. Data science is a field that relates to data cleansing, preparation and analysis. Data science algorithms are used in many industries like Internet searches, Digital Advertisements, Travelling, Healthcare, Gaming, Financial services etc. There are various applications in today's world where data analysis and data visualization is required. Data science can solve the problems like classification, identifying anomalies, to quantify, finding way of organization, decision making issues etc. In this paper, we have shown how python is useful and acts as a key to solve such problems. In addition to python, there are also some other platforms which are used to solve a task completely based on data science. Here we have focused on python and it's packages that are highly useful for data science based problems. We have shown how python can be used for data analysis and data visualization.

Key words: Data Science, Python, Data Analysis, Data Visualization

I. INTRODUCTION

Data science is highly attractive field due to the reason that it can solve decision making problems through the data itself. In fact, it's much more than this. Usually it solves the problems as:

- Classification based problems like to identify the outcome among some fixed values.
- To detect anomalies or outliers
- Regression based problems like how much temperature will be on Monday?
- Clustering based problems to understand structure of data. e.g. to identify the viewers who like same movie.
- What should I do next? i.e. decision making problems.

All the above problems are industrial based and is impressively helpful for any organization to look forward and build their organization in a strong position. Data science answers sharp questions. If an organization is interested in identifying the stock's sale price for the next week, it can solved by using data science. Data science may not help in answering vague questions. Questions should be sharp enough that is answerable with data.

The complete Data science is not relevant to a single person only. It's a team that involves statisticians, computer scientists, AI scientists and experts in other relevant fields. Data science by definition is interdisciplinary and requires not just method disciplines but also requires the domain science [1]. Data science comprises many areas like Machine Learning and optimization, Mathematics, Statistics, Information theory, Information technology. Data involved in the process may vary according to the disciplines. Scientific data is used in bioinformatics, social informatics etc. On the other hand if business data is there, then companies like Amazon, eBay, Google, facebook etc. use the data to build or predict new business strategy. Amazon uses collaborative filtering to recommend high quality product to the customers and facebook use people you may know feature to recommend friend connections .

In this paper, we will discuss how python can be used for the solution of all mentioned problems of data science. There are many tools and technologies that are used for data science. One of the highly or effectively used programming language for data science is python. It acts as a key for data science. There are many features because of which python is highly useful for data science. We will discuss the interesting features and packages such as pandas, numpy, matplotlib etc. that make python very useful and effective for data science. We will also discuss the features that are involved from basic to advanced level of python and is useful for data science.

II. STEPS IN DATA SCIENCE

The following are the steps involved in a data science process:

- Acquire data
- Preparation of data
- Analysis of data
- Preparation of report
- Apply results or Act on results

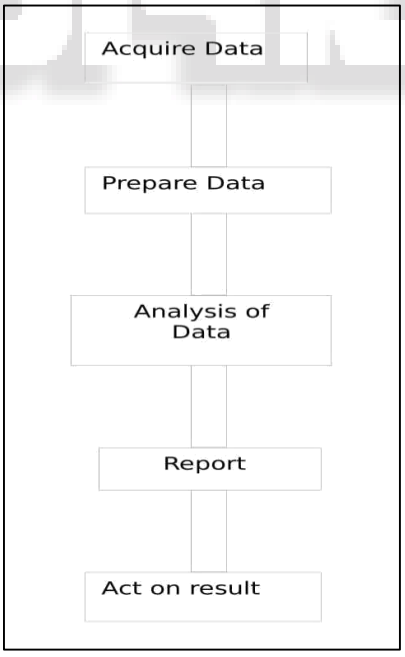

Fig. 1: Steps involved in data science process

A. *Acquire data:*

It means identifying data sets and retrieve data that is required for the problem. Data should be real to get fine results. Data collection can be done by different ways e.g data may be available online, or it may be collected from outside world. But it should be relevant to the problem which we try to solve. (Identify data sets, Retrieve data, Query data)

B. Prepare data:

Once the data is acquired, then data is prepared for analysis. This is most important step to be performed. As much time we spend on this part, we get as better result. In preparation of data, we explore the data by doing preliminary analysis and then understanding nature of data. Then we preprocess the data by cleaning and integrating. The received data may not be perfect as there may be missing values, redundancy, unusable features so preprocessing involves solving all such problems e.g. missing values may be filled according to the feature. It may be average of all values available, or it may be mode of the feature or something else. So, this step of data science is important from results point of view. To get better result, better preprocessing is required. Pandas library provides fast and flexible data structure that can be used to work with relational and labeled datasets in an easy way [2]. As preprocessing of the data set takes 70% of the time to make data set useful and according to as per requirement of algorithms. The goal of data preprocessing is to choose cardinal features then remove irrelevant information and finally transform raw data into sessions[3] .

C. Analyze data:

(Select analytical techniques, Build models): Now, once the data is prepared then some analytical technique is applied on the data to build a model. A model is built so that data may be visualized in graphical form and most importantly it can be analyzed. Analysis involves looking for the result on the basis of certain features.

D. Preparation of report:

Once result is analyzed, a report is prepared on the basis of features that the relevant to the problem. Using this analytical report, some actions or decisions can be made for the future point of view.

E. Act:

As soon as report is prepared, action takes place. If the result found is decisive, then decision making is done as per the relevant problem. Result may be in the form of classification, clustering, reinforcement etc. so that future decision can be made on the basis of analysis. The complete process is an iterative process.

III. USE OF PYTHON

Python programming language is highly useful for data science. As data analysis, data visualization are key parts of data science and these tasks can be effectively solved by using python. There are many libraries in python that can be used to solve such tasks. In fact, from data preparation to data visualization, python programming language can be used. Here we will discuss about some of the libraries and their uses.

As data science revolves around data analysis and data visualization. Python has wonderful libraries for these tasks. There is NumPy library that can be used to analyze the data using various numerical data analysis techniques . Another library is SciPy that can be used for statistical tests. Similarly Matplotlib library can be used to produce graphs for the analysis done . As we know data preparation is very important for producing an effective model. Python is used for the data preparation also. As we saw earlier that cleansing, integration etc. are the parts of preparation so such tasks can be easily done by python. So now we have seen from data preparation to data visualization, all tasks can be performed using python programming language. Scripts in Python programming language have more advantages than models in the software like ModelBuilder. Basic program structures like loop (for, while), decision making statement (if, else) can be utilized. Moreover, there is high probability to avoid errors during a program run
[5].

There is one more library pandas that is available with python. This is the library of rich data structures and tools for working with structured data sets common to statistics, finance, social sciences, and many other fields. Apart from these useful libraries,feature selection and extraction are the important concerns [6].

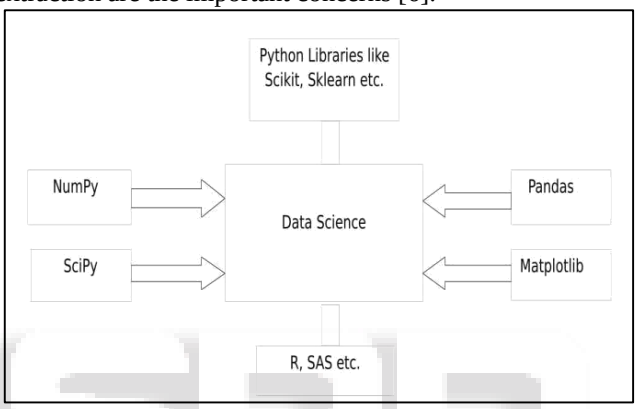

Fig. 2: Python Libraries with other skills required for data science

MongoDB database can be used to store the dataset that will be used for applying data science processes. So the above discussed libraries are as follows :

- Pandas (for data manipulation and analysis)
- NumPy (for analysis)
- SciPy (for statistical tests)
- Matplotlib (for graphical presentation)

These libraries have many of the methods that are used to solve some specific tasks. So these libraries are very useful and powerful for doing data science in python. Python also has other types of libraries that can be used for some computing purpose e.g the Python library multiprocessing is a package that supports spawning subprocesses to leverage multicore shared-memory resources [4].

IV. MACHINE LEARNING USING PYTHON

Machine Learning is the field where an algorithm is produced on the basis of data. Data comprises of training data and testing data. Using training data, model is preapared and then using that model, testing data is applied on the model and performance or accuracy is checked. There are various machine learning algorithms such as supervised algorithms, unsupervised algorithms and reinforcement algorithms. Supervised algorithms are used for classification and regression. Unsupervised algorithms are used for clustering.

When testing data is applied on the model prepared through training data and machine learning techniques, then we get a result. Such results are analyzed and visualized to

present the final form of analysis. Using visualizations, decisions are taken by company or organizations to improve their benefits and all.

V. SKILLS REQUIRED FOR DATA SCIENCE

We have seen the importance of python for data science. It can be used for preprocessing, analysis, statistical tests, visualization etc. In addition to command on python, there are some more skills which a data science should have. A data scientist should know how to work with unstructured data. Normally a developer can easily work with structured data using SQL database but to work with unstructured data, Hadoop platform can be used. A data scientist requires mathematical skills, analytical skills etc. Other than these, a data scientist should also know SAS and R . However, a data science is not done by a single person in any organization. There is a specific team for that but the mentioned skills are required for being a data scientist. If we analyze the things then we can see that it's like artificial intelligence where we can get prediction on the basis of data. Human-Level Artificial Intelligence is the intelligence of a (hypothetical) machine that could successfully perform any intellectual task that a human being can [7].

VI. CONCLUSION

We have seen the data science process which involves raw data collection, data preparation, data analysis, communicate result and act on the result. So from data collection to data visualization, python programming language is highly powerful and useful. We have discussed different libraries of python for different kinds of tasks such as Pandas for data manipulation and analysis, Numpy for analyasis, Scipy for statistical tests, Matplotlib for graphical presentation of result. Apart from these libraries, there are many more libraries available in python that are effective for data science. There are some other languages and tools also available that are used for different purpose.

Like we discussed earlier, SAS and R are also used in data science. R language can be used for data preprocessing but the same can be done by using python also. So we have seen that many of the tasks related to data science can be performed using python programming language. In this way, we can say that python acts as a key programming language for data science. We can conclude that the Python is flexible language and provide suitable tools to perform data sciences techniques and running a [8].

REFERENCES

[1] Javin D. West, "The science of data science", Journal of Integrated creative studies, No. 2016-010-e, May 2016.
[2] Wes McKinney, "pandas: a Foundational Python Library for DataAnalysis and Statistics", DLR Portal,www.dlr.de/sc/Portaldata/15/Resources/dokument e/.../pyhpc2011_submission_9.pdf.
[3] Gaurav, Zunaid Alam, "Road Safety in india using Data Mining Approach", International Conference, REDSET 2017, CCIS 799, https://doi.org/10.1007/978-981-10-8527-7_17
[4] Rosa Filguiera, Iraklis Klampanos, Amrey Krause, Mario David, Alexander Moreno and Malcolm Atkinson, "A Python Framework for Data-Intensive Scientific Computing", IEEE Conference, 978-1-4673-6750-9, Nov-2014.
[5] Ing. Zdena Dobesova, "Programming Language Python for Data Processing", IEEE Conference, 978-1-4244-8165-1/11, Sept-2011.
[6] Fabien Dubosson , Stefano Bromuri, and Michael Schumacher, "A Python Framework for Exhaustive Machine Learning Algorithms and Features Evaluations ", IEEE Conference, 1550-445X/16, March-2016.
[7] Ankur Bhatia, "Artificial Intelligence – Making an Intelligent personal assistant", International Journal of Computer Science and Engineering, Vol. 6, No. 6, 2015.
[8] Nurul Afiqah Mat Zaib , Nor Erne Nazira Bazin ,Noorfa Haszlinna Mustaffa , Roselina Sallehuddin, "Integration of System Dynamics with Big Data Using Python: An Overview", IEEE Conference, 978-1-5090-6255-3/17, May 2017

