International Journal of New Technology and Research (IJNTR)
ISSN: 2454-4116, Volume-7, Issue-6, June 2021 Pages 71-75

Python for Data Science

Shalini Khandelwal, Vishwas Dhabria, Ankit Tiwari

***Abstract*— Python is an object-oriented programming language that is gaining popularity in the field of data science and analytics by building sophisticated software applications. Python has the largest and most powerful libraries used to analyze and visualize data. Data scientists have to deal with a lot of data known as big data. With easy use and a large set of python libraries, Python has become a popular option for managing large amounts of data. Python builds better analytical tools that can help data scientists in building machine learning models, web resources, data mining, segmentation etc. In this paper we will review the various tools used by python editors to make data analytics more efficient and comprehensive and compare with other languages.**

***Index Terms*— Machine learning, data science, big data.**

I. INTRODUCTION

Highlight The fast-growing digital information is moving faster over the internet and much of it has random information namely photos, video, sounds, blogs, tweets, facebook posts, Google map location and much more. Traditional methods of handling such complex data that are less organized are particularly challenging in the software industry [1]. For applications such as big data, data science, social media analytics and market research software specialist uses python which provides a standardized library for methodical machine learning and data analysis.

II. DATA ANALYTICS AND ITS LIFE CYCLE

Data analytics is a great method to analyze data to bring out the meaningful insight from the data.

A. ***Requirement Understanding:-*** In this section we need to understand the basic requirement for analysis such as why it is needed, its use and basic information. The process can be long and difficult so we need a road map to do the same.

B. ***Data Collection:-*** In this section, various data sources are available depending on the type &size of the problem. Many data resources mean more opportunities to find hidden connections and patterns. Keyword capture tools, data and information from these different data sources are needed. Organized and unplanned data required should be stored in a database. NoSQL database information is required to enter Big Data. Various frameworks and websites are created by organizations such as Apache, Oracle etc. which allows analytics tools to download and process data from these repositories.

C. ***Data Cleaning:-*** This section is dedicated to the removal of duplicate, corrupt, useless and unrelated data items. This section uses business-based verification rules to verify the need and value of the data extracted for testing. Although it can sometimes be difficult to apply verification issues to the extracted data due to difficulty. Integration helps to combine multiple data sets into smaller numbers depending on common fields. This facilitates data processing.

D. ***Data Analysis and Processing:-***This section enables real-time data mining and analysis to establish unique and hidden patterns for business decisions. The process of data analytics can vary depending on the situation i.e., testing, verification, guessing, determined, diagnostic or descriptive.

E. ***Result Interpretation:*** **-**This section includes the representation of the results of the analysis into a visual or graphic form that makes it easy for the audience to understand.

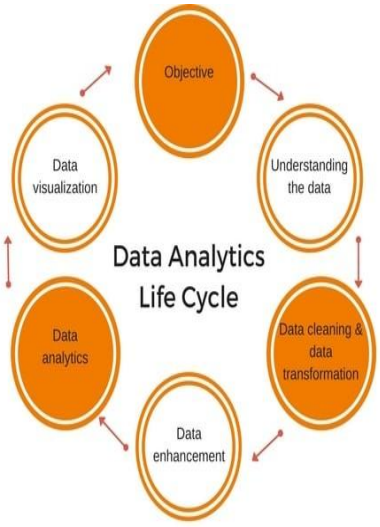

Fig 1 Cycle of Data Analysis Process

III. DATA MINING TOOLS

Prior to data mining, data must be cleaned and processed from its raw state. With various similar tools available such as Microsoft Excel and Google Spreadsheets. The downside to these tools is that they can't be used for big data sets, even though they can be used for small-scale engineering [2]

A. EDM

EDM is a tool used for automatic distillation and data

Shalini Khandelwal, Computer Science &Engineering, Vivekananda Institute of Technology, Jaipur, India
Vishwas Dhabria, Computer Science &Engineering, Vivekananda Institute of Technology , Jaipur,India
Ankit Tiwari, Assistant Professor, Dept. of Computer Science & Engineering, Vivekananda Institute of Technology, Jaipur, India

label. Much of the distillation of the EDM workbench feature is attributed to certain excel errors and Spreadsheets for specific data-related tasks, such as the production of complex sequence features, data sampling, labeling, and data integration into user-friendly subsets that teach students according to user-defined terms. [2]. The work bench now allows scientists to study [3]

A pre-collected educational data label with interesting behavioral categories (e.g. playing the system, helping to avoid), is much faster than possible with previous live viewing or existing data labeling methods.

• Collaborate with others in writing data.

• Automatically extract additional information from log files for use in machine learning, such as estimating student information and content about student response time (e.g. how fast or slower student action is than the measure of that problem step).

B. SQL

SQL, or Structured Query Language, is used to organize specific information. SQL queries can be a powerful way to extract the information you want, sometimes integrating ("joining") into multiple data tables. [2]

SQL can be used by a data scientist for basic analysis such as generating query from the query, managing dates, digging documents, finding moderators, uploading data to your database and creating sequences. [4]

IV. DATA MINING ALGORITHMS FOR DATAANALYSIS

Once the features are well-designed and well-organized, we need some expertise to record the data collected for continuous analysis and feature prediction.

A. Rapid Miner:- RapidMiner can upload and analyze any type of data including both formal and informal such as text, images and media. It has access to more than 40 types of files including SAS, ARFF, Stata and URL. It provides support for all data including NOSQL, MangoDB, Oracle, IBM DB2, Microsoft SQL Server, Postgres, Teradata, Ingres, Vector Wise and many more. It also allows access to cloud storage like Dropbox and Amazon3. [5]

B. WEKA:- Waikato Environment for Knowledge Analysis is a free and open-sourcesoftware package that integrates a wide variety of data mining and construction model algorithms. Keep has a comprehensive set of organizational, collection, and cohesive mining techniques that can be used alone or in combination, in ways such as bag placement, power booster and stack installation. Users can request algorithms for extracting data from the command line, GUI (graphical interface), or through the Java API. Keep can produce the models they produce depending on the realitymathematical models, or in PMML files (Predictive Modeling Markup Language) can be used to model new data using the Apply score plugin to use the model. [3]

C. KNIME:-KNIME ("naim", KoNstanz Information MinEr, www.knime.org), formerly Hades, is a data purification and analysis package that offers a wide range of specialized algorithms in areas such as emotional analysis and social network analysis. The most powerful feature of KNIME is its ability to combine data from multiple sources (e.g. .csv of built-in features, text document name response, and student statistics database) into a similar analysis. KNIME also offers extensions that allow it to integrate with R, Python, Java and SQL. [3]

D. KEEL:- KEEL, which is open sourcea software tool available under GNU to assess the evolutionary potential of Data Mining problems of various types including retransmission, fragmentation, uncontrolled reading, etc. It incorporates evolutionary algorithms based on a variety of methods: Pittsburgh, Michigan and IRL, as well as the integration of evolutionary learning strategies with different pre-processing strategies, allowing it to perform a comprehensive analysis of any learning model in comparison with existing software tools. [6]

E. Tableau:- Tableau is an easy-to-use tool for creating a structured, interactive interface. Although Tableau is often discussed in the context of business intelligence, it can also be used to create a positive scientific and creative perspective in the context of research, public health, and medical care. Although Tableau's drag-and-drop interface is more user-friendly and readable than most other visual tools, the efficient use of software requires some practice, as well as good practice habits in data viewing. [7]

R Language:- R is an open source data analysis and programming language. R contains a number of mathematical algorithms for machine learning and machine learning that allow users to perform retrospective research and develop data products. R has a diverse, open and free software that assists in big data processing. R has the ability to integrate with many other programming languages such as C ++, Java. It can store items on a hard disk and process them intelligently. [8].

V. PYTHON FOR DATA ANALYSIS

Python features make it a perfect match for easy-to-read, solid, readable, awesome data, a comprehensive set of libraries, integration with other languages and an active community and support system. Python Libraries for Data Analysis [9]:

Table 1: Python Libraries and its Functions

Library	Usage
Numpy, Scipy	Statistical and technical computing
Pandas	Data manipulation and aggregation
Mlpy,scikit-learn	Machine Learning

Theano, tensor flow, keras	Deep learning
Stats models	Statistical analysis
Nltk, genism	Text processing
Network x	Network analysis and visualization
Bokeh,matplotlib,seaborn, plotly	Visualization
Beautiful soup, scrapy	Web scraping

A. The Top 5 Development Environments

Python provide different editors for different applications.But there are a few other editors that can be used in the field of data science. [11]

Spyder: - In contrast to most of the IDE's available on the web, Spyder was designed specifically for data science. It includes features such as text editor with syntax highlighting, code completion, and variable exploring that can be changed with the help of the Graphical User Interface (GUI).

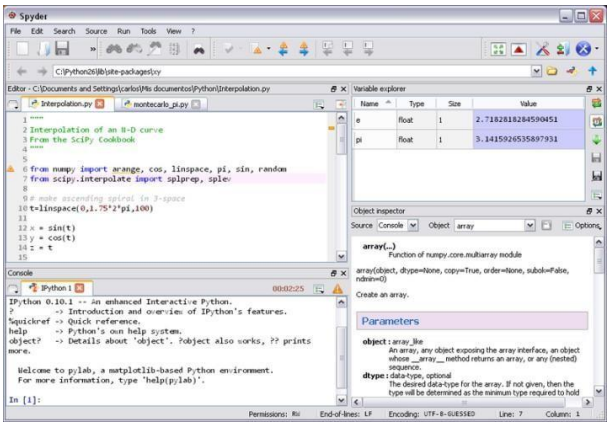

Fig.2SpyderIDE

Py Charm: - PyCharm is an IDE created by the people behind Jet Brain, the team responsible for one of Java's most popular IDE, the IntelliJ IDEA. PyCharm integrates its tools and libraries like NumPy and Matplotlib, allowing you to work with multiple viewers and interactive devices. Other than Python, PyCharm supports JavaScript, HTML/CSS, Angular JS, Node.js, and many more, which highlights it as a great option for web developers. Much like other IDEs, PyCharm has interesting features such as code editor, errors highlighter, a powerful debugger with a graphical interface, in addition to the integration of Git, SVN, and Mercurial. You can customize your IDE, choose between certain themes, color schemes, and key bindings. Moreover, you can expand PyCharm's features by adding certain plugins.

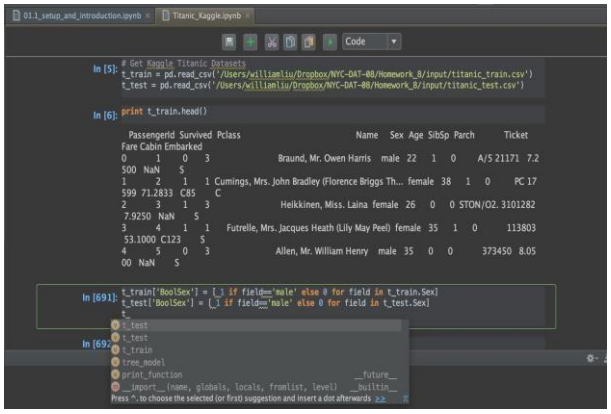

Fig.3 Py Charm IDE

Thonny: -The next in this sequence is Thonny an IDE suitable for learning and teaching programming. It has been at The University of Tartu. Defining its features, Thonny supports code completion and highlighting syntax errors, and it also provides a simple debugger, which can run your program step-by-step. This has to be the most suitable for beginners, as they can step through statements and expressions. In the process of editing a function, a new window is created with local variables and the code is presented separately from the main code.

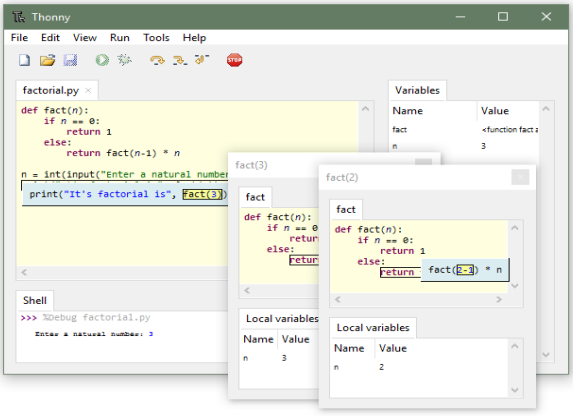

Fig.4 Thonny IDE

: - An open-source text editor developed by GitHub. With this editor being available for many programming languages, such as Ruby on Rails, PHP, Java, etc., Atom has features that makes it a great experience for Python developers. The highlighting advantage of Atom is its developer community, particularly as a result of improvements in the field of tools and plug-ins that are designed to adapt to your IDE and use it to improve your workflow.

For example, one of these plug-ins, which is termed as "Packets", is the Data Atom, allowing you to write and run SQL queries. It supports PostgreSQL, Microsoft SQL Server, and MySQL. Additionally, you can use the "Visualize your Result" tool of Atom, allowing you to verify the final output without the need of opening another window. There is also a plug-in called 'Markdown Preview Plus", which offers built-in support for editing and previewing Markdown files, which you can use to open, view, and visualize the LaTeX equations. And, much like any other IDE, you can use a

variety of panels, patterns, and colors, helping you in managing multiple projects.

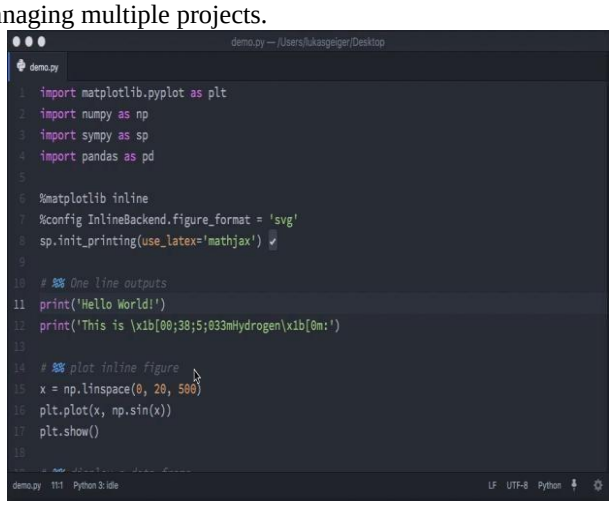

Fig.5Atom IDE

Jupyter: - Jupyter Notebook that was born from IPython in 2014, is a web-application which is based on a client-server framework, allowing you to create and edit notebooks documents, or simply the "notebooks". Jupyter Notebook offers an easy-to-use, interactive, data-science environment for many programming languages, and that it works not just as an IDE but as a presentation or a training tool as well. This is also an ideal place to be for those who are just stepped into data science. Jupyter Notepad supports markdown, so that you can add HTML elements from images to videos. Apart from that you can make use of data visualization libraries such as Matplotlib and Seaborn, that shows graphs n the same document that contains the code. In addition, you can export your work as PDF or HTML files, or simply export them as .py. Also, you can make blogs and presentations from your notebook.

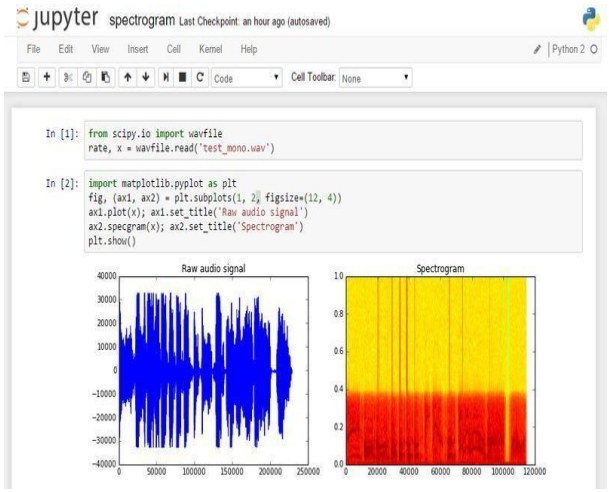

Fig.6JupyterIDE

VI. CONCLUSION

Python provides a variety of libraries and editors for efficient data analysis. Python is the fastest-growing language that is being extensively used by data scientist for analytics purposes, such as You tube, Google, and others. In addition to the mathematical research that Python supports, a wide variety of computing resources are in the hand of those who are well-versed in the language. The research team is particularly interested in the development of algorithms for the modern and distributed supercomputers that use GPUs to accelerate computing. As you can see, Python is an efficient tool for cutting-edge research in data science. Of course, there are a lot of such tools, and often the specific choice of the language used in the data study it is a matter of personal taste. However, it is respectfully asserted that some of the languages have a wide-range of support for the data science research than what Python has to offer.

REFERENCES

[1] Randy Paffenroth, Xiangnan Kong, Proc. Of the 14th Python in Science Conf. (SCIPY 2015) https://www.youtube.com/watch?v=EUEHOYl0mR "Python in Data Science Research and Education"

[2] Slater, S., Joksimovic, S., Kovanovic, V., Baker, R.S., Gasevic, D. "Tools for educational data mining: a review"

[3] Rodrigo, M. M. T., Baker, R.S.j.D., McLaren, B.M., Jayme, A. & Dy, T.T. (2012). In: K. Yacef, O. Zaïane, H. Hershkovitz, M. Yudelson, & J. Stamper, J. (Eds.) Proceedings of the 5th International Conference on Educational Data Mining (EDM 2012). (pp. 152-155) "Development of a Workbench to Address the Educational Data Mining Bottleneck"

[4] https://www.mastersindatascience.org/data-scientist- skills/sql/

[5] https://rapidminer.com/products/studio/feature-list/

[6] J. Alcal´a-Fdez1, L. S´anchez2, S. Garc´ıa1, M.J. del Jesus3, S. Ventura4, J.M. Garrell5, J. Otero2, C. Romero4, J. Bacardit6, V.M. Rivas3, J.C. Fern´andez4, F. Herrera1 "KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems"

[7] Federer, Lisa M., and Douglas J. Joubert. 2018. "Providing Library Support for Interactive Scientific and Biomedical Visualizations with Tableau." Journal of eScience Librarianship 7(1): e1120. https://doi.org/10.7191/jeslib.2018.1120

[8] Sanchita Patil MCA Department, Vivekanand Education Society's Institute of Technology, Chembur, Mumbai in International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056 Volume: 03 Issue: 07 | July-2016 www.irjet.net p-ISSN: 2395-0072 © 2016, IRJET | Impact Factor value: 4.45 | ISO 9001:2008 Certified Journal | Page 78 "Big Data Analytics Using R"

[9] Kang P.Lee ITS-RS/U13 "Introduction to Python Data Analytics" in June 2017 10, RP1

[10] https://www.datacamp.com/community/tutorials/data-science-python-ide

[11] Thirunavukkarasu K1 and Dr.Manoj Wadhawa in International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.6, No.1, February 2016 "Analysis and Comparisison Study of Data Minimg Algorithms Using Rapidminer".

[12] Makrufa Sh. Hajirahimova, Marziya I. Ismayilova DOI: 10.25045/jpit. v09.i1.07 Institute of Information Technology of ANAS, Baku, Azerbaijan "Big Data Visualization: Existing Approaches and Problems".

[13] Ms. Komal, International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES) Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2585 Volume 4, Issue 5, May-2018 IJTIMES-2018@All rights reserved 1012 "A Review Paper on Big Data Analytics Tools"

[14] Kalpana Rangra Dr. K. L. Bansal, Volume 4, Issue 6, June 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering Research Paper (ijarcsse) "Comparative Study of Data Mining Tools"

[15] Anmol Bansal and Dr. Satyajee Srivastava et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 5, Issue 1, March 2018, pp. 15-18 "Tools Used in Data Analysis: A Comparative Study"

[16] Ken Kelly, Keke Lai and Po-Ju Wu, A Best Practice for Research "Using R for Data Analysis"

[17] Dr. Snezhana Sulova, Dr. Latinka Todoranova, Dr. Bonimir Penchev, Radka Nacheva, Bulgaria www.sgem.org "Using Text Mining to Classify Research Papers"

[18] D G Rossiter Version 1.4; May 6, 2017 "An example of statistical data analysis using the R environment for statistical computing"

[19] K. R. Srinath Telangana, India International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 04 Issue: 12 | Dec-2017 www.irjet.net p-ISSN: 2395-0072 © 2017, IRJET | Impact Factor value: 6.171 Page 354 "Python – The Fastest Growing Programming Language"

[20] Shivangi Kaushal Jagpuneet Kaur Bajwa,Mohali India International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 10, October 2012 ISSN: 2277 128X www.ijarcsse.com "Analytical Review of User Perceived Testing Techniques"

[21] Nawsher Khan, Ibrar Yaqoob,Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed KamaleldinMahmoud Ali,Muhammad Alam,, Muhammad Shiraz and Abdullah Gani, Malaysia Hindawi Volume 2014, Article ID 712826, 18 pages http://dx.doi.org/10.1155/2014/712826 "Big Data: Survey, Technologies, Opportunities, and Challenges"