iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35683

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

Python: A Quintessential approach towards Data Science

Aniket M. Wazarkar
Student, Computer Engineering Department, Savitribai Phule Pune University

Abstract: ***Python is an interpreted object-oriented programming language that is sustainably procuring vogue in the field of data science and analytics by fabricating complex software applications. Establishing a righteous nexus between developers and data scientists. Python has undoubtedly become paramount for data scientists mindful of cosmic and robust standard libraries which are used for analyzing and visualizing the data. Data scientists have to deal with the exceedingly large amount of data alias as big data. With elementary usage and a vast set of python libraries, Python has doubtlessly become an admired option to handle big data. Python has developed and evolved analytical tools which can help data scientist in developing machine learning models, web services, data mining, data classification, exploratory data analysis, etc. In this paper, we will scrutinize various tools which are used by python programmers for efficient data analytics, their scope with contrast to other programming languages.***

Keywords: Python, Data Science, Machine Learning, Big Data, Deep Learning .

I. INTRODUCTION

The digital information (data) is growing at brisk pace over the internet as per current scenarios are concerned, and the upcoming brisk in the future is unparalleled. The major section of this what we called as unstructured data comprises of images, audio, video, Facebook, Instagram, Twitter posts, map locations, financial transactions and much more. So, handling this unstructured data the traditional way is a big ultimatum for the software and data industry. This is where python comes into play, for such entities like big data, data science, data engineering, business analytics and market research Python simplifies the approach of software professionals, providing the dynamic standard libraries for orderly machine learning and data analytics. The language constructs ensure the user to write clear programs on all scales whether it is small body or large body. Python supports multiple programming paradigms including imperative, object-oriented, and functional programming or procedural making it more and more effective. Automatic memory management is enabled by default.

II. LIFE CYCLE OF DATA ANALYTICS

Data analytics is the process and methodology of analyzing data to draw meaningful insight from the data. Provided that the data given may be unstructured or structured. So, there involves some series of steps to make sure data is properly processed for visualization.

1) *Requirement Understanding:* It basically is the specific requirement guidelines like what we need for the process i.e. application and basic details. The process is long, lengthy and hence we must progress with a road map to do the same..
2) *Data Collection:* Talking about this phase, here the wide variety of data sources are identified depending upon the severity of the problem. As the number of data resources surges so do the chances of finding hidden correlations and patterns. While working on Big Data, tools and frameworks are very much need to capture data from heterogeneous data sources. Thus, the partially processed data needs to be stored in the database. NoSQL databases are needed to accommodate Big Data, MongoDB can be preferred best for business as per the condition. Organizations like Apache, Oracle, etc. have developed adequate frameworks that assist analytics tools for fetching and processing data from these repositories.
3) *Data Cleaning:* This phase plays an important role in the removal or cleansing of replicated, null, irrelevant data objects from the dataset or data source. This stage applies validation rules based on the test case to ensure the necessity and relevance of extracted data for analysis. Although, due to the complexity of data sometimes it becomes almost impossible to apply validation constrained to the extracted data. Aggregation groups values from multiple entities together, making the dataset simplified for further processing.
4) *Data Analysis and Processing:* The actual data mining and analysis is been carried out in this stage to establish unique and hidden patterns for making business decisions. Provided from many data analysis techniques i.e. exploratory, confirmatory, predictive, prescriptive, diagnostic, or descriptive, the technique may vary depending upon the business point of view.
5) *Result Interpretation:* In this phase the analytical result is represented in visual or graphical form, making it easy for the audience to read and understand.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

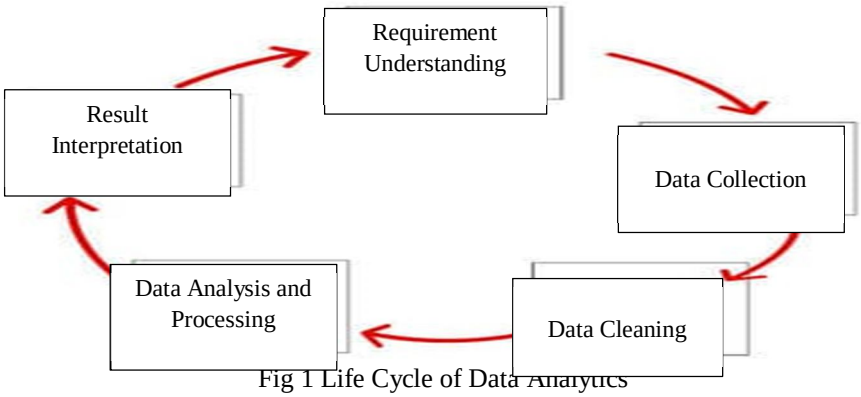


Fig 1 Life Cycle of Data Analytics

III. TOOL FOR DATA MINING

Before data mining, the data must be cleaned and processed from their raw state. This can be done by some tools, some of the best of them are listed below,

1) *Monkey Learn:* Having a user-friendly interface, MonkeyLearn is easy to integrate with your existing tools in order to perform real-time data mining. Having pre-trained text mining models like sentiment analyzer or to build a custom solution to cater to more specific business needs MonkeyLearn has it all.
2) *Oracle Data Mining:* It is a component of Oracle Advanced Analytics. It enables data analysts to build and implement predictive models. It contains several data mining algorithms for classification, regression, anomaly detection, prediction, and more, making it super productive for Data Scientists, Data Miners.
3) *Rapid Miner:* The ability of RapidMiner to load and analyze any type of data including both structured and unstructured text, images, and media makes it simply outstanding. Having access to more than 40 file types including SAS, ARFF, Stata also supporting via URL makes it best for business. RapidMiner has got support to almost all databases including NOSQL, MangoDB, Oracle, IBM DB2, Microsoft SQL Server, Postgres, Teradata, Ingres, VectorWise, and many more. RapidMiner also has the ability to access cloud like Amazon and Dropbox making it easier for professionals to work remotely.
4) *Weka:* It is an open-source machine learning software having a vast collection of built-in algorithms for data mining. It was developed by the University of Waikato, in New Zealand. It provides support for different data mining tasks, like preprocessing, classification, regression, clustering, and visualization. Having a good graphical interface makes it easier to use.
5) *Orange*: Orange is a free, open-source data science toolbox. It operates in sections of the development, testing, and visualizing data mining workflows. Being component-based software, it has a large collection of built-in machine learning algorithms and text mining add-ons. It also provides some extended functionalities to bioinformaticians and molecular biologists. It offers numerous graphics like silhouette plots and sieve diagrams. Orange's drag and drop facility helps non-programmers to perform data mining tasks through virtual programming. However, developers may opt to mine data in Python.
6) *Tableau:* One of the easiest to use tools for creating customized and interactive visualizations. Tableau's main functionality includes creating effective scientific and biomedical visualizations in the context of research, public health, and medical care. Its drag-and-drop interface is easy to use, user-friendly, and easy to learn as compared to other data visualization tools. Although it does require some familiarity with best practices in data visualization.

IV. DATA MINING ALGORITHMS FOR DATA ANALYSIS

Data mining algorithms are being mostly used in Data Science, Artificial Intelligence, and Machine learning. Python is generally used to implement these algorithms.

1) *C4.5 Algorithm:* One of the best data mining algorithms, developed by Ross Quinlan. It is a supervised learning algorithm. It generates a classifier in the form of a decision tree from a dataset. However, the dataset needs to be classified. Classifier means a basic data mining tool that is used to take data as an input which we need to classify and then predict the class of new data. Decision trees are very facile to interpret which makes C4.5 a popular and fast algorithm as compared to other data mining algorithms.
2) *SVM:* SVM, also known as Support Vector Machine. It works similarly that of C4.5 algorithm discussed above, except for the fact that SVM does not make use of any decision tree. However, the hyperplane concept is used to classify data into two classes. The hyperplane equation of a line can be mathematically visualized as "y=mx+b". The main ascendancy of SVM is that it can project our data to higher dimensions.

3) *Apriori Algorithm:* Being an unsupervised learning algorithm, it is used for discovering new, interesting patterns and find the relationships among the dataset. The working mechanism of Apriori algorithm is based on association rules. Although the algorithm is highly efficient, it consumes an exorbitant amount of memory, disk space and is time-consuming.
4) *K-mean Algorithm:* Undoubtedly, one of the most popular clustering algorithms, k-means functions by creating k number of clusters based on the data set based on the similarity of objects in the dataset. K-mean does not guarantee the group members will be exactly similar, but it ensures that the group members are more similar as compared to non-group members. Being an unsupervised learning algorithm, it learns about the cluster on its own without any external interference.
5) *CART Algorithm:* CART, also known as classification and regression trees. It is nothing but a decision tree-based algorithm which gives either regression or classification tree as an output. Like C4.5, CART is also a classifier. The regression or classification tree model is prepared by using the labeled training data set provided by the user itself, making it a supervised learning algorithm.
6) *kNN Algorithm:* kNN is also known as k-Nearest Neighbor is used as a classification algorithm. However, it is been considered a lazy learning algorithm. Unlike C4.5, SVM which are eager learners i.e. they start to build the model during training itself, kNN will classify only when it receives new unlabeled data as an input.

V. PYTHON FOR DATA ANALYSIS

The extensive features of python make it perfect for data analytics i.e. easy to learn, robust, readable, scalable, ability to integrate with other programming languages, extensive set of libraries, and vast community and support system.

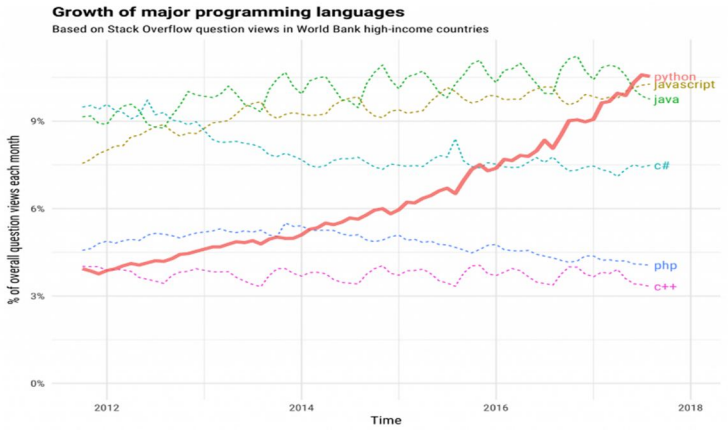


Fig.1 Growth of Python as a Programming language over time.

Some of the most likely and popular Python libraries for data analytics are:

Libraries	Usage/Function
NumPy	It is used for performing complex mathematical operations like Fourier transformation, linear algebra, random number etc.
SciPy	SciPy provides statistics, optimizations, integration and linear algebra packages for advanced computation.
PANDAS	It is used for high-performance data structures and analysis.
Matplotlib	It is used for 2D plotting like histograms, plots, bar charts, scatter plots etc.
SciKit-Learn	It is being used for classification, regression and clustering.
Theano	It is used for fast numerical computation that can be run on the CPU or GPU
Keras	It is used for high-level neural networks APIs for integration
PyTorch	It is used for deep learning applications which uses CPU's and GPU's and dynamic computational graph design.
PyBrain	It provides complex yet powerful machine learning algorithms.

Tenserflow	It is used to create large scale neural networks.
Seaborn	It is used to visualize the complex statistical models
Bokeh	It is used for creating interactive visualizations for modern web browsers
Plotly	It is used to create beautiful interactive web-based visualizations that can be displayed in Jupyter notebooks.
NLTK	It is used to perform operations like tagging, tokenization, stemming, semantic reasoning etc.
Gensim	It is used for topic modeling and space vector computations
Scrapy	It is used to collect data through APIs and act like a crawler
Statsmodels	It provides providing the data exploration modules to perform statistical analysis and assertion.
Kivy	It provides a neutral user interface which allows to develop multi-platform applications.
PyQt	It provides platform-independent abstractions for Graphical User Interfaces, SQL databases, OpenGL, XML.
OpenCV	It provides huge library for computer vision, machine learning and image processing which helps in processing images and videos to identify objects, faces, or even handwriting of a human and much more.

Table 1: Python Libraries and Corresponding Functions

VI. LEADING PYTHON IDE'S FOR DATA SCIENCE

Python provides different editors as per the approach of the professionals. Some of the topnotch Python editors which can be used specifically for data science are listed below:

1) *Jupyter Notebook*: It is a free open-source, interactive, user-friendly, web tool also known as a computation notebook. Here the data science professionals can be used to combine software resources and tools, exploratory text and data, computational output, and multimedia resources in a single file. It provides an easy-to-use environment, across many data science programming languages which do not work only as an IDE, but also as a presentation, education, and exploratory tool. One who wants to begin his journey in data science can begin with jupyter notebook. We can add HTML components from images to videos in Jupyter Notebook using its markdown function. Almost all the Python libraries can be imported in jupyter, e.g. NumPy, SciPy, Matplotlib, Seaborn, etc. Besides all these functionalities one can export his final work to formats like PDF, HTML, ipynb, txt, .Py, etc.

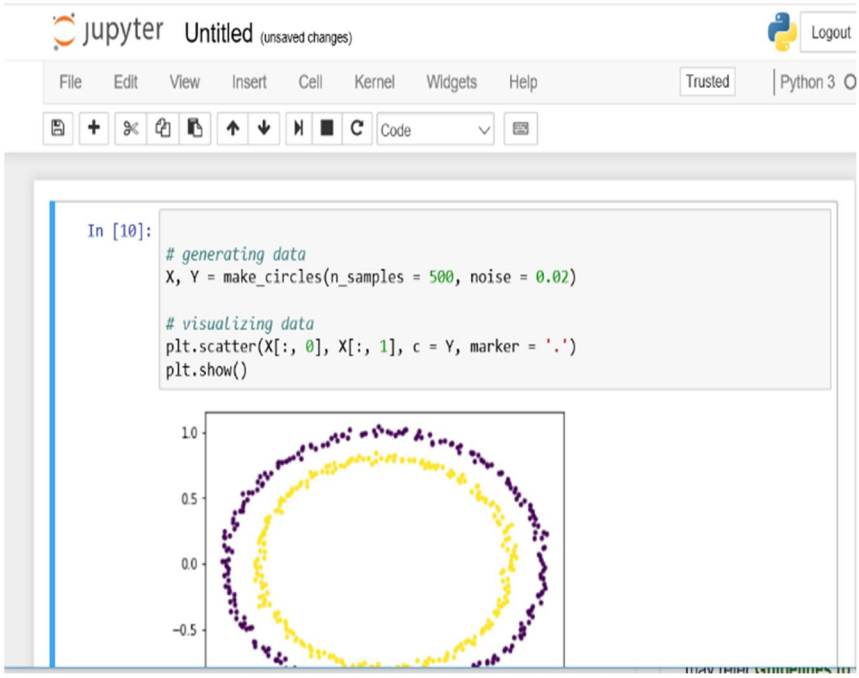


Fig. 2 Jupyter IDE.

2) *PyCharm:* The PyCharm is developed by a company called JetBrains, formerly known as IntelliJ. PyCharm is undoubtedly one of the most popular Python IDE, being a cross-platform software, it supports Windows, Linux as well as Mac. Pycharm provides intelligent code compilation along with easy project navigation, on-the-fly error debugging, linter integration, and much more. It duly supports Python's web frameworks like Flask, Django, Web2Py, etc. Pycharm has integrated a unified interface with version control systems like Git, perforce, Mercurial, etc. It also provides an API with the help of which the user can develop their own plugins to extend PyCharm features. In addition, with Python, has got support for JavaScript, HTML, AngularJS, and many more.

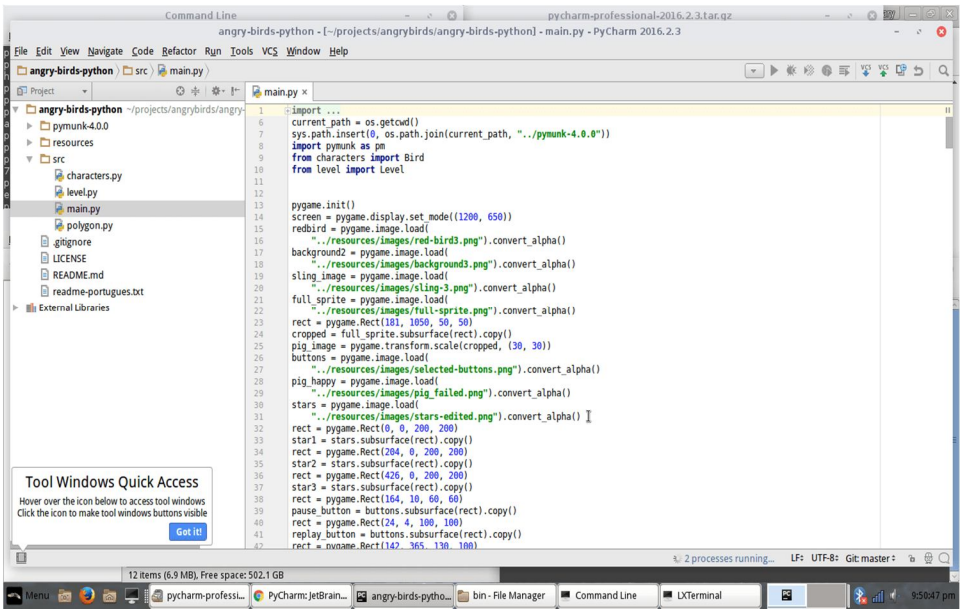


Fig. 3 PyCharm IDE

3) *Spyder:* Spyder is a cross-platform, open-source scientific IDE for Python designed especially for data science. It provides features like smart code compilation, syntax highlighter, and introspection, i.e. it examines the property of an object at runtime. The professional edition of Spyder does provide support for Python's web frameworks like Flask, Django, Web2Py. Also, it has support for scientific tools like NumPy, SciPy, Matplotlib, Seaborn, etc. Having the version control integration with version control systems like Git, Mercurial, Perforce, CVC results in ease for the developers to commit and revert changes accordingly.

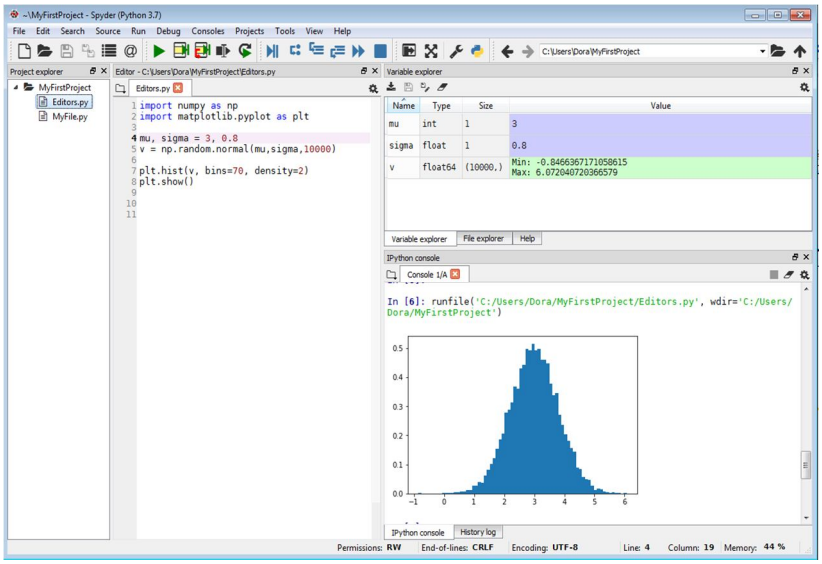


Fig. 4 Spyder IDE

4) *Thonny:* Developed by The University of Tartu, Thonny is a cross-platform, open-source IDE. Talking about the features, it is a simple debugger where one can execute the program step by step. It is a good and beginner-friendly debugger. One can easily understand how Python works under the hood scoop.

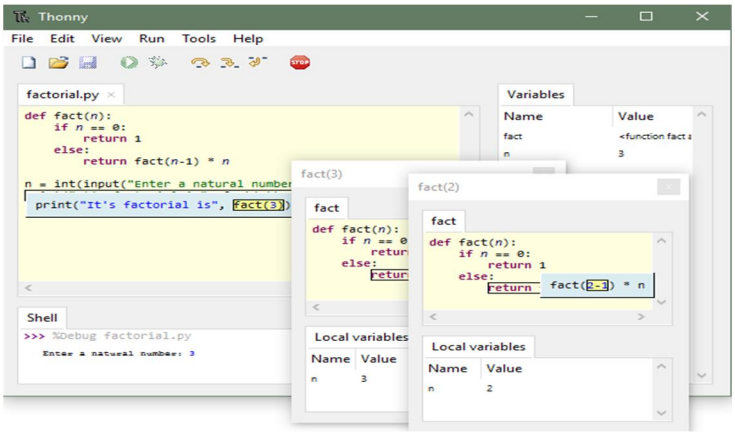


Fig. 5 Thonny IDE

5) *Atom:* Atom is an open-source software created by GitHub. Atom, being a cross-platform software provides its support to Windows, Linux as well as macOS. Atom was written in CoffeeScript and Less, but most of it is been converted into JavaScript later on. Being fully customizable in HTML, CSS, and JavaScript, developers call it a "hackable text editor for the 21st century". It is purely based on Electron (previously known as Atom Shell), making it cross-platform desktop application. It supports syntax highlighting.

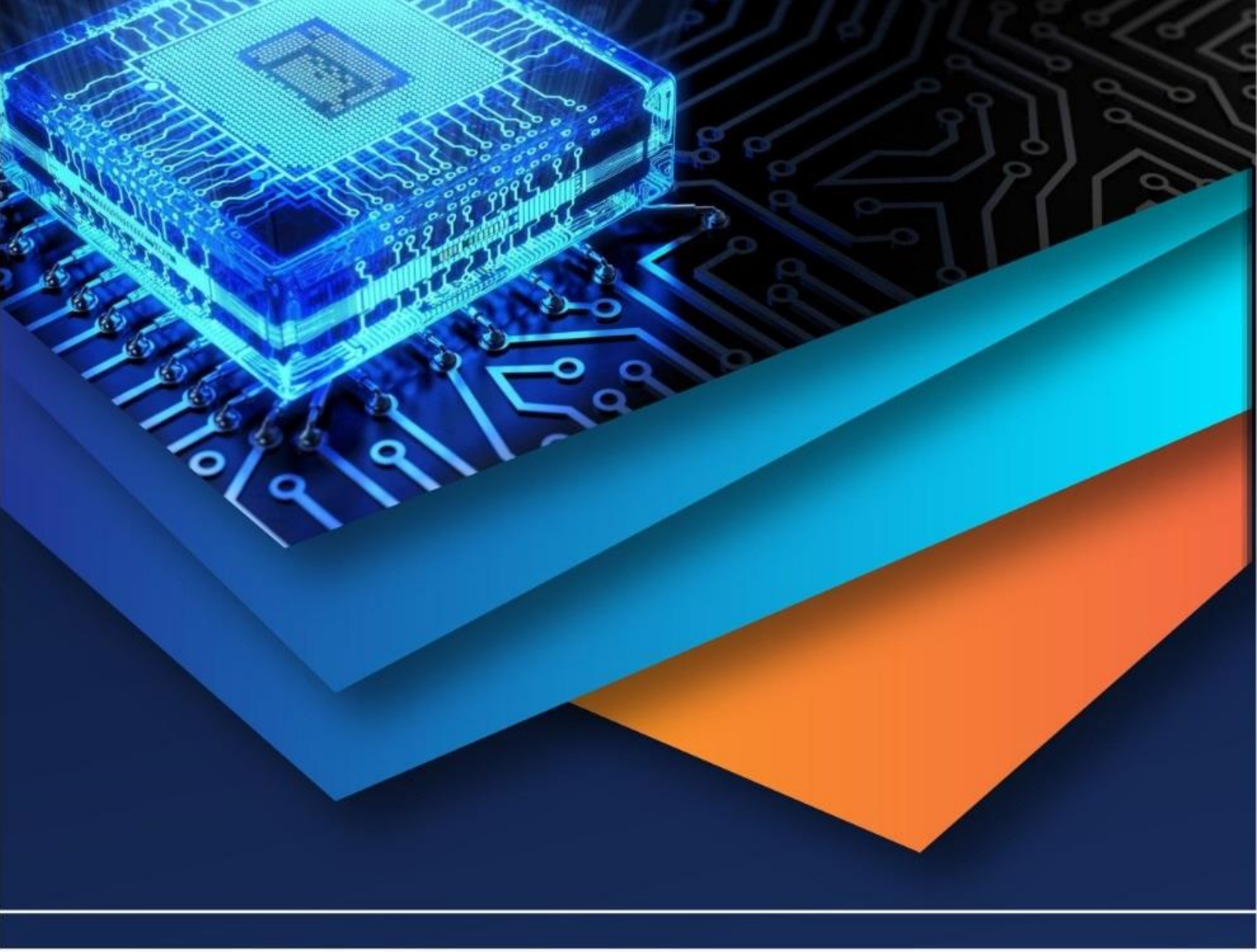
Fig. 6 Atom IDE

VII. CONCLUSION

Apart from being flexible, robust, easy to learn, Python is one of the fastest-growing languages, which means there is huge active community support available. Python is the foremost valuable tool of Data Scientists. It is made for repetitive tasks and data manipulations. Only the one who has worked on the large datasets knows how many time repetitions come into play. Almost all the Tech-Giants like Google, Facebook, Amazon, IBM are using Python for data science and analytics. Beyond the mathematical computation that Python supports, there are a vast array of computational resources that are at the fingertips of the one who is a connoisseur in Python. The Python research groups are always engaged in advancing the algorithms for modern distributed supercomputers that would leverage GPUs to accelerate computations. Therefore, Python appears to be perfect from all points of view for cutting-edge data science research and development. There are many tools, languages for data science research purposes which is nothing but a matter of taste and flexibility. However, Python being flexible, robust, precise, and having vast global tech support makes it unparalleled, cutting-edge technology for "Data Science".

REFERENCES

[1] Analytics and Data Science Standardization and Assessment Framework. (2020). Harvard Data Science Review. doi:10.1162/99608f92.1a99e67a
[2] Pierson, L., & Porway, J. (2017). Data science. Hoboken, NJ: John Wiley & Sons.
[3] Pierson L, Porway J (2017) Data science. John Wiley & Sons, Inc., Hoboken, NJ
[4] Rogel-Salazar, Jesús. "Python: For Something Completely Different." Data Science and Analytics with Python, 2018, pp. 31–85., doi:10.1201/9781315151670-2.
[5] Asay, Matt. "Python Is Devouring Data Science." InfoWorld, InfoWorld, 20 Apr. 2021, www.infoworld.com/article/3615695/it-s-pythons-all-the-way-down.html.
[6] Manaranjan Pradhan, U Dinesh Kumar (2020). Machine Learning using Python, Wiley India Pvt. Ltd.
[7] Schutt, R., & ONeil, C. (2013). Doing data science. Beijing: OReilly Media.
[8] Hasija Y, Chakraborty R (2021) Python for Data Visualization. Hands-On Data Science for Biologists Using Python 91–122. doi: 10.1201/9781003090113-5-5.
[9] Science Behind Social Media Data. (2013), Bloomberg.
[10] Newland, S. (2021). Machine learning. London: Wayland.
[11] Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice. Cambridge, MA: Morgan Kaufmann.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY