World Journal of Research and Review (WJRR)
ISSN: 2455-3956, Volume-14, Issue-5, May 2022 Pages 15-19

Review Paper on Python for Data Science & Web Development

Rajkumar Chouhan, Kapil Singh, Ritu Vashistha

***Abstract*—Reasons for using python in the data science: 1.Easy to code, 2.Free and open source3.Scalability, 4. Libraries and frame work like: Numpy, Pandas,Scipy and Matplotlib. 5. Huge Community, 6.Web Development Framework of Python is Django, Flask and Pyramid. So above reasons explain how python has become a popular option to Handle Data science and Web. Python builds better analytics and visualizing tools which can help in developing many machine learning models, web services, data mining & classification. In this review paper we will know about the tools which used in field of data science.**

***Index Terms*— Python libraries, Data science, Machine learning & Web Framework.**

I. INTRODUCTION

Python programing use in the field of data science, IoT, AI, and other technologies. This language contains costly tools from a mathematical or statistical perspective. In the today internet life data is growth rapidly and the major portion is to handle the unstructured data i.e. video, images, blogs, fb posts and applications of Google. Handling of this unstructured data is not easy task. This is a big challenge to software industry. In this python language help the developers or data scientist with their powerful libraries. Python provides efficient tools for visualizing and understanding complex problems [1, 2].

II. PYTHON LIBARAIES

1) Numpy:- This library was created in 2005 by Travis Oliphant which stands for Numerical Python. This library consisting of multidimensional array objects and collections of paths for processing those arrays. Using this library we can perform mathematical and logical operations on the array. This is also called linear algebra library. Numpy statistical functions are Mean, Median, Standard deviation, Variance, Average and Percentile etc.
   Numpy array operations are checking array dimensions, array indexing, array creations, indexing and slicing and advanced methods on arrays. There are various features of this numpy which is high performance, integrating code from C/C++, multidimensional container, work with varied databases, broadcasting functions and additional linear algebra [2, 3].

2) Pandas: - This library is created by the Wes McKinney in 2008. The name panda has a reference to both “Panel Data” and “Python Data Analysis”. This library is used for working with data sets. It has function for the cleaning, analyzing, exploring, and manipulating data. There are many features of pandas library which are multiple file formats supported, great handling of data, handling missing data, grouping, merging and joining data sets alignment and indexing and visualize. To read CSV(comma separated values)file, the read_csv() method of the Pandas library is used. The major application of pandas is in Economics, Recommendation system, Stock prediction, Neuroscience, Statistics, and Natural Language Processing.

3) Matplotlib: - It was introduced in 2002 by John Hunter. Matplotlib is the most popular data visualization library in python. It allows to create figures and plots, and makes it very easy to produce static raster or vector files without the need of any GUI’s. This is a multi- platform data visualization library built on Numpy and also use as making 2D plots of arrays.
   There are two approaches of creating plots which is Functional approach and Object-oriented interface. It provide an object-oriented API for embedding plots into applications using general purpose GUI toolkits like Tkinter, wxPython, Qt, and GTK.

4) Scipy: - Scipy stands for the Scientific Python. This is scientific computation library uses Numpy underneath. It provides more utility function for stats, optimization, and signal processing. This was created by Numpy’s creator Travis Oliphant. Few segments of this library are written in C language. This library mainly efficient in calculation of linear algebra, integration, ordinary differential equation solving and signal processing.

5) Scikit-learn: - This was initially developed by David Cournapeau as a Google summer code project in 2007. It is most powerful library for machine learning and data science. It provides statistical modeling including regression, clustering, and dimensionality reduction via a consistence interface in Python. The fundamental base of this library is

Rajkumar Chouhan, B.Tech Scholar, Vivekananda Institute of Technology, Jaipur
Kapil Singh, B.Tech Scholar, Vivekananda Institute of Technology, Jaipur
Ritu Vashistha, Assistant Professor, Department of CSE, Vivekananda Institute of Technology, Jaipur

Numpy, Matplotlib, and Scipy [4, 5].

III. DATA SCIENCE

Data Science is an interdisciplinary field that uses scientific method, processes, algorithm and system to extract to knowledge and insights form structured and unstructured data. In simple words we can say it combines domain of programming skills, expertise, and knowledge of math and statistics to extract meaningful insights from the data [6, 7].

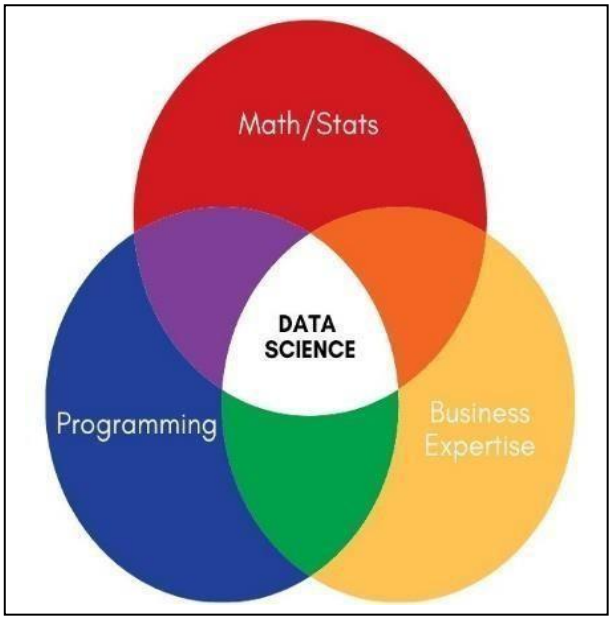


Fig. 1: Data Science

A. Role of Data Science in Business Expertise

Data science method can explore historical, make comparisons to competition, analyze the market, and make recommendationsof about your product sells. Data scientist is responsible for advising companies on data potential, gaining new sights and transforming them into business goals. Developing the solutions using data science technologies like Statistical analysis, Data mining, Data visualization for improve business performance. Deep knowledge of data science concepts and tools can enable you to make more understand your business performance and plan for future.

B. Data Mining

Data mining is technique to extract important information from a huge set of data. We can also say it's about finding the trends in data set. It often includes analyzing the vast amount of historical data which was previously ignored. This is subset of data science. There are several applications of data mining which are Future Healthcare, Market Basket Analysis, Fraud Detection, Financial Banking.

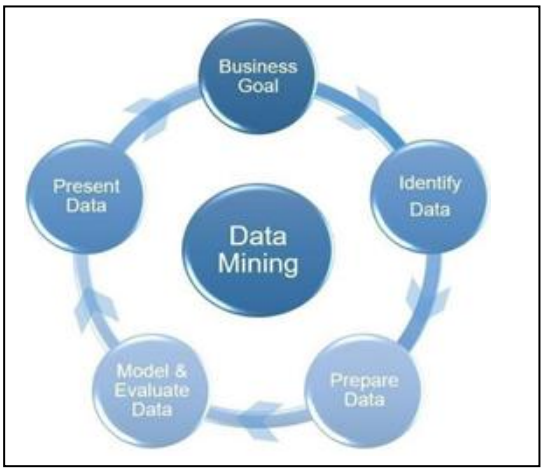


Fig. 2: Data Mining

C. Data Science Life Cycle

Data Science life cycle is a process of deliver the data science project or product. Every specific data science life cycle is different because every project and teams are different. For a successful life cycle, it is important to understand each and every section well and distinguish all the different parts. You have also a good knowledge of Development stage versus the Deployment stage, as they have different requirement that need to be satisfied as well the business aspect.

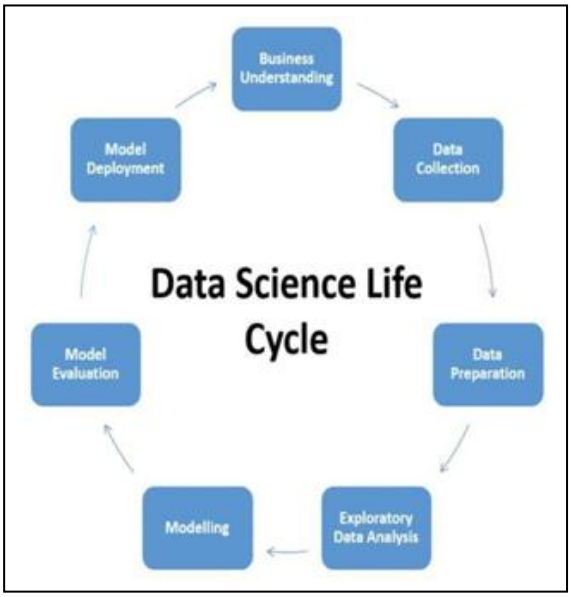


Fig. 3: C. Data Science Life Cycle

D. Need of Data Science

Data Science is vital in almost every field. It needs to develop and progress within its system to handle emerging issues in every industry, business, and organization. Data science comes into picture where problem and tasks are the advance level. The principal purpose of data science is to find patterns within the data by various type of statistical techniques. Data science is the need of multiple fields like Marketing, Customer Acquisition, Innovation & Enriching lives etc. [8].

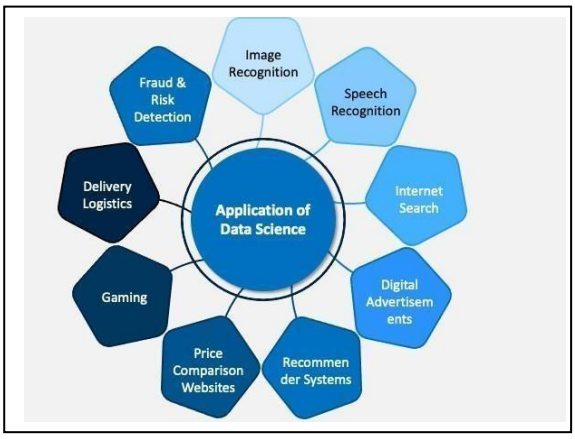


Fig. 4: Application of Data Science

IV. MACHINE LEARNING

The term Machine leaning is introduced by Arthur Samuel in the year of 1959. Machine Learning provides ability in system to learn automatically and improve from experience without being explicitly programmed. It is an application of artificial intelligence. Learning process starts with observation or data like direct experience, in order to look for data and make better decisions in future based on the examples that provide [9, 10, 11].

Types of Machine learning

1) Supervised Learning: - Supervised learning is the process of making an algorithm to learn to map an input to a particular output. In this learning our data set is labeled which means well defined. Correct mapping makes better the algorithm and our model predict perfectly. This learning can help make predictions for new unseen data that we obtain later in the future.

   There are two types of supervised learning Regression and Classification. Regression algorithm learns from the labeled datasets and then able to predict a continuous-valued output for the new data given to the algorithm. In classification techniques classes need to be mapped to either 1 or 0 which in real life translated to 'Yes' or 'No'. This type of learning is use in Bioinformatics, Speech Recognition, Spam Detection, and Object-Recognition for Vision.

2) Unsupervised Learning: - In this learning data is unlabeled and you are unsure about the outputs. It helps in modeling, probability density functions, finding anomalies in the data, and much more.

   There are two types of Unsupervised learning Clustering and Association. In Clustering learning you find patterns in the data that you are working on. It may be the shape, size, color etc. In Association learning you find the dependencies of one data item to another data item and map such that they help you profit better. This type of learning is use in Amazon, AirBnB, and Credit-card fraud detection.

3) Reinforcement Learning: - This is known as agent based learning. In this an agent learns to behave in an environment by performing actions and seeing the result of actions. The agent learns automatically sing feedback.

   There are two types of this learning which is Positive leaning and Negative learning. In Positive learning when an event occurs due to a particular behavior, increase the strength and the frequency of the behavior. In Negative learning strengthening of behavior because a negative condition is avoided. This type of leaning has the practical application like robotics in industrial automation, data processing, creating training system.

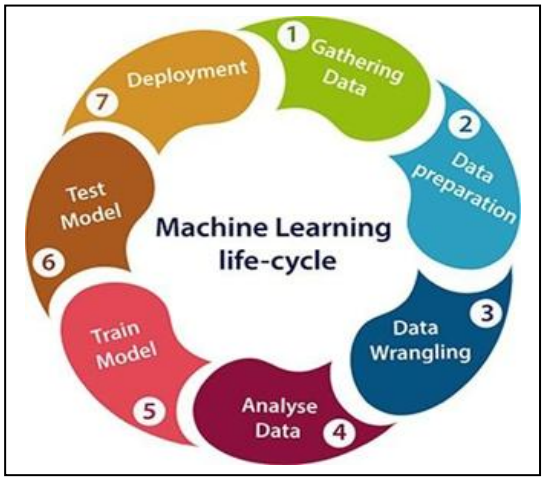
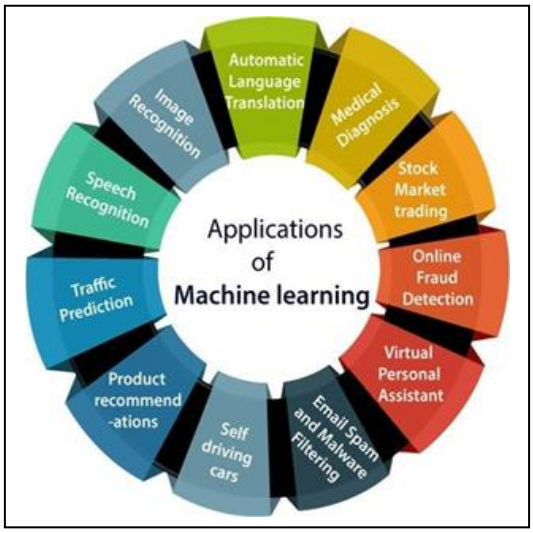


Fig. 5: Machine Learning Life-cycle

Need and Application of ML:

Machine learning provides smart alternatives to analyzing the large volume of data. By developing fast and efficient algorithm and data-driven models for real-time processing of data, Machine learning can produce accurate results and analysis. It makes computer get into self-learning mode without explicit programming. When fed new data, these computers learn, grow, change & develop by themselves. ML allows to organization to transform the large data sets into knowledgeable and action intelligence. This information can be integrated into everyday business process and operational activities to respond to changing market demands or market circumstances.

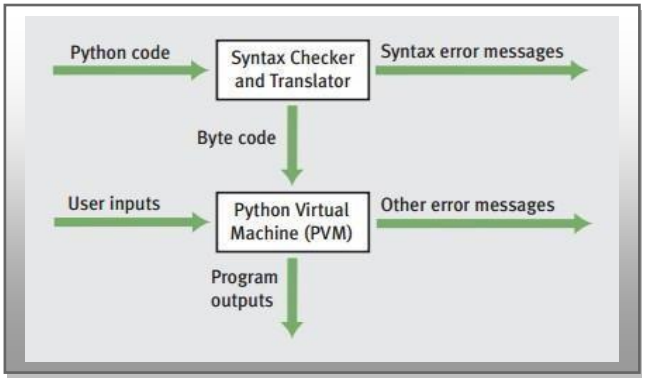


V. DEVELOPMENT ENVIRONMETS

A development environment is a combination of text editor and a python runtime implementation. The text editor allows you to write code for your applications. The runtime implementation is the method of executing your code. Python provides many editors for different application. We will study about the most important editor which is use in data science [12, 13].

1. ***PyCharm:-*** This is developed by the Czech Company JetBrains(Known as IntelliJ). It is an integrated development environment (IDE) use for python programming language. In pycharm, you can see where and how symbols, such as tags, classes, fields, methods, or functions are defined in you project. For this purpose, the IDE features the Quick definition popup. This IDE has many attractive features like setting up virtual environment, downloading python modules, amazing code completion.
2. ***Jupyter Notebook: -*** This is created by Fernando Perez in 2012. Jupyter is open source web application that you can use to create and share documents that contain live code, visualization, equations, and text. It supports the Python, R, Julia programming language. This app is server-client application that allows editing and running notebook via a web browser. Major features of Jupyter notebook is the ability to display plots that are the output of running code cells, very easy to host sever side, which is useful for security purpose, and useful for other people's work as a starting point.
3. ***Spyder:*** - It is created by Pierre Raybaut in 2009, since 2012 Spyder has been maintained. This in an open-source cross platform IDE that include with Anaconda. It includes editing, interactive testing, debugging, and introspection features. Spyder is written in same language python that you use it to develop, so it's easy to get started contributing guide.
4. ***Atom: -*** Atom is a free and open source text and source code editor developed by GitHub. This is for Linux, Windows, and macOS with support for plug-ins written in JavaScript, and embedded GitHub control and also known as "hackable text editor for the 21st Century".

VI. INTERNAL WORKING OF PYTHON



The standard implementation of python is called "cpython" and we use c codes to get output in python. Python converts the source code into a series of byte codes. Python code is translated into intermediate code, which has to be executed by a virtual machine, known as Python Virtual Machine (PVM).

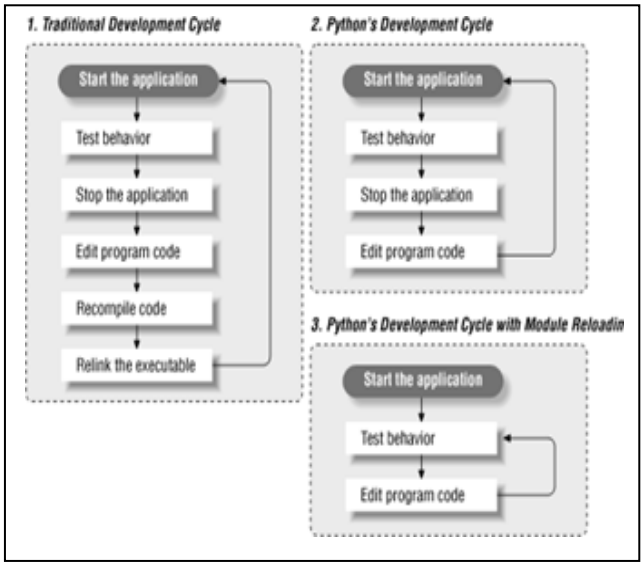


Fig. 7: Python Development Cycle

VII. WEB DEVELOPMENT WITH PYTHON

In Web Development Python include sending data to and from Servers, Processing data and Communicating with Database URL Routing and ensuring security.
Django, Flask and Pyramid are the Main Python Framework that's used in web development [14, 15, and 16].

1. Django

Django was design and developed by Lawrence journal world in 2003 and publicly released under BSD license in July 2005.
Currently, DSF (Django Software Foundation) maintains its development and release cycle. Django was released on 21, July 2005.
Django is a high-level Python web framework that enables rapid development of secure and maintainable websites. Built by experienced developers, Django takes care of much of the hassle of web development, so you can focus on writing your app without needing to reinvent the wheel. It is free and open source, has a thriving and active community, great documentation, and many options for free and paid-for support. Django is based on MVT (Model-View-Template) architecture. MVT is a software design pattern for developing a web application. View: The View is the user interface — what you see in your browser when you render a website. It is represented by HTML/CSS/Javascript and Jinja files.

Model: The Model is the part of the web-app which acts as a mediator between the website interface and the database. In technical terms, it is the object which implements the logic for the application's data domain. There are times when the application may only take data in a particular dataset, and directly send it to the view (UI component) without needing any database then the dataset is considered as a model.

View: View is actually the User Interface of the web-application and contains the parts like HTML, CSS and other frontend technologies. Generally, this UI creates from the Models component, i.e., the content comes from the Models component.

Controller: The controller as the name suggests is the main control component. What that means is, the controller handles the user interaction and selects a view according to the model.
The main task of the controller is to select a view component according to the user interaction and also applying the model component. This architecture has lots of advantages and that's why Django is also based on this architecture. It takes the same model to an advanced level.

MTV Pattern: Django is mainly an MTV (Model-Template-View) framework. It uses the terminology Templates for Views and Views for Controller. Template relates to the View in the MVC pattern as it refers to the presentation layer that manages the presentation logic in the framework and essentially controls the content to display and how to display it for the user.

VIII. CONCLUSION

In this paper we have discussed about the use of python in Data science field. We also know that how python is efficient language for the Machine learning models. Developer uses python because of its easy syntax and fast working. Python has more number of libraries which is use in solving complex algorithms, various calculation tasks and model building. Python also plays the important role in data pipelines automation and works fast in compare to the Excel. Python can accomplish most day-to-day research tasks and can be used at multiple steps of research pipeline (e.g. running experiments with participants, data organization, data processing/manipulation, statistical analysis/modeling and visualization).

These frameworks make web application development a very sophisticated and organized procedure and help build scalable and efficient applications. Also, they have the ability to hold both client-side and server-side programming contents. So, we can conclude that Python language is the future and Tester will have to upgrade the skills and learn this language to tame the AI, ML, Data Science and Web Development.

REFERENCES

[1] Python Data Science Handbook: Essential Tools for working with data by Jake VanderPlas.
[2] https://www.researchgate.net/publication/335874810_Python_in_Field_of_Data_Science_A_Review
[3] https://towardsdatascience.com/tagged/python
[4] https://www.datacamp.com/community/tutorials/data-science-python-ide
[5] Understanding Theory of Machine Learning from Theory to Algorithms by Shai Shalev -Shwartz and Shai Ben-David.
[6] https://docs.python.org/3/library/
[7]
[8] https://www.javatpoint.com/data-science
[9] Giullio Rossetti, Letizia Milli, Fosca Giannotti International Journal of Data Science and Analytics 5,61-79 (2018), "A python library to model and analyze diffusion processes over complex networks".
[10] Sarkar, Dipanjan, Bali, Raghav Sharma, Tushar "Practical machine learning with python" A problem- solver's guide to Building a Real-World intelligent system.
[11] Cheng,X., Jing,W., Song,X., Lu,Z. 5th International Conference of Pioneering Computer Scientist, Engineering and Educators, ICPCSEE2019, Guilin, China, September 20-23,2019,Proceeding, Part 1"Data Science".
[12] https://www.edureka.co/blog/introduction-to- machine-learning/
[13] https://www.tutorialspoint.com/python_data_science/ index.htm
[14] Arabnia, H.R.,Daimi K.,Stahlbock,R.,Soviany,C., Heillig, L., Brussau, K.©2020 "Principle of Data Science".
[15] https://analyticsindiamag.com/a-complete-tour-of-data-science-project-life-cycle/
[16] https://www.sketchbubble.com/en/presentation-application-of-data-science.html