International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 6 | ISSN : 2456-3307

# Guidance to Data Mining in Python

**Aashish Mamgain**
CSE, HMR Institute of Technology and Management, New Delhi, India

**ABSTRACT**

Python has become top programming language in the field of data mining in recent years. Around 45% of data scientists are using python programming language for data mining. Python is ahead from other analytical tools such as R. Data mining is the technique in which large datasets is analyzed for generating predictive patterns, information. Data mining is used to detect various applications such as marketing, medical, telecommunications and so on. This paper presents classification algorithms such as Random Forest, Support Vector Machine, Decision Tree, Logistic Regression etc. This guide provides data mining classification techniques in python programming language.

**Keywords :** Python, Data Mining, Classification, Random Forest, Support Vector Machine, Decision Tree, Logistic Regression

## I. INTRODUCTION

Python has become popular in data analysis because it is easy to grasp. Its syntax is easy to understand and readable. Operating on large dataset, commonly known as Big Data. More data has to process, the more important it, because to manage the memory used. That's why python is used by data scientists.

Python has libraries for data analysis visualizations, statistics, natural language processing. Python libraries for data analysis includes Scikit-learn, Numpy, Pandas, Scipy, Tensor Flow and so on [1]. This immense library provides data scientists with a huge amount of structural and fuctionality. Python can be interacted directly with code or using terminal or other tools like Jupyter notebook, this is the advantage of python, other than any language. Data mining and Machine learning are subset of Artificial Intelligence and are both related process, in which the data is analyzed. It is essential for these processes to have tools that allow quick iteration and easy interaction.

In Data mining, there are three approaches which includes classification, regression and clustering. Classification is dependent on supervised learning. It most popular algorithm in data mining. Classification is not only used to study and examine the existing sample data but also predicts the future behaviour to that sample data. The classification has two stages, one is learning stage in which the training datasets is analysed, then the rules and patterns or information are generated. The second stage tests or examine the datasets and achieve the accuracy of classification patterns or information generated in stage one. Clustering is dependently based on unsupervised learning because there are no pre-existing classes, that is there is no labelled data. In this approach, unlabelled data or entities is grouped together, knowns as a cluster. Regression is used to map input data with the output data and regression is used for prediction of values. In this paper, we will discuss python libraries and various machine learning algorithms includes neural network, svm, decision tree, random forest etc.

The paper has been segmented as follow: In section 1 presents brief introduction about paper. In section 2, brief introduction python libraries used for data mining. In section 3, discuss different machine learning algorithms used for deep learning. In section 4, there is comparative study related to algorithms and programming language. In section 5, gives conclusion derived from this article.

## II. Python Libraries

Python helps data scientists in order to analyse data mining by proving numerous libraries. These are as follow:

### A. Sci-kit learn Library

scikit-learn provides algorithms for machine learning including classification, regression, dimensionality reduction, and clustering [2]. Scikit-learn is open source library and tackle well-known machine learning algorithms by providing robustness. It range between supervised learning and unsupervised learning[3]. It is easy-to-use and designed to integrate with python scientific library.

### B. Numpy Library

NumPy is the library for computing scientific calculations with Python. NumPy package specializes in multi-dimensional arrays processing in which arrays performs sophisticate functions and grant element-by-element operations [4]. It has various features including use of linear algebra, random number, fourier series and tool for integrated with C/C++. This takes out the worries that usually mire quick programming in other languages.

### C. Pandas Library

Pandas is package for python programming which is used for data analysis. It is easy-to-use and provide high performance. Panda library is used for indexing, data manipulation of dataframe. It provides functions for reading and writing data in different formats such as CSV, text files, sql database etc.

### D. Scipy Library

SciPy is a Python library and an open-source software for computing scientific calculation. It is integrated on the numpy object [4] and provides mathematical functions like integrations, special functions, optimizations and so on to new level in scientific programming. Sometimes, numpy framework also called scipy frameworks.

### E. Matplotlib Library

Matplotlib is a Python library which produce 2D charts, histograms, heat maps and so on various figures. Matplotlib is used in Python IDLE, Jupyter notebook and various tools used in data analysis. It provides functional API's to plot graphs into applications. Matplotlib has pyplot module for plotting graphs which provides interface.

### F. Tensorflow Library

Tensorflow is open source python package for computing numerical data. It is developed by Google Brain to perform artificial intelligence research. It is used in neural networks applications. It is easy to implement on application because of its flexible architecture. User can easy build models from scratch because of its contented classes and functions.

## III. Machine Learning Algorithms

Machine learning algorithms are divided into three categories namely- Supervised learning, Unsupervised learning and Reinforcement learning. The main techniques are discuss below.

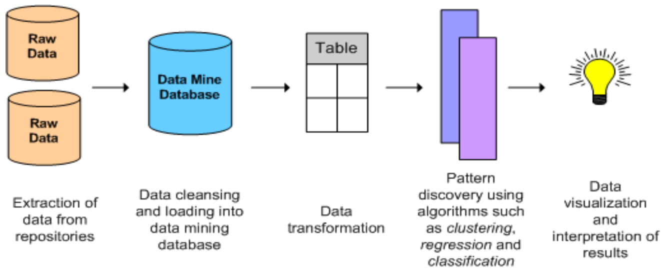
Raw Data
Raw Data
Data Mine Database
Table
Extraction of data from repositories
Data cleansing and loading into data mining database
Data transformation
Pattern discovery using algorithms such as *clustering, regression* and *classification*
Data visualization and interpretation of results

**Figure 1.** Block Diagram of Data Mining Process

### A. Random Forest

Random Forest is a supervised learning algorithm. RF is a learning method used for classification and regression. Generally, the more number of trees in the forest means more robustness forest. Random forest works on strategy, more the number of tress, higher will be the accuracy result. It is collection of tree structure classifiers in which tree is generated using training sets for class input x [11]. It accuracy is good and sometimes even better. It dependency lies between classifier individually.

### B. Decision Tree

Decision tree is categorized under supervised learning that is commonly used in classification tasks. Basically, it works on continuous input and output elements. In this technique, each node represents instance to be classified. Each leaf represents class label. It generates rule for classification techniques. There are basically three algorithms namely- CART, C4.5 and ID3 [12].

### C. K-means clustering

K-means clustering is undertaken by unsupervised learning, which is used the data is unlabelled and that unlabelled data is grouped together, commonly known as clustering. The main focus of this algorithm is to make cluster of data. These groups can be represented by the variable K [13]. K-means works on two phases. The first phase is set K centre randomly, where K is fixed. The second phase is to take each data to the nearest centre [14]. Generally, data points are used in global clustering.

### D. Neural Network

It is categorized under unsupervised learning. Donald Hebb's proposed examples of unsupervised learning. He states that neurons which fire together are wire together. It approaches to problem solving tasks rather than conventional computers [15]. Due to its remarkable properties have many applications like image processing, character recognition, forecasting and many more. They cannot be programmed to perform specific tasks. It main aim to perform tasks by composing large number of processing elements (neurons).

### E. Support Vector Machine

SVM is undertaken by supervised learning. It is basically training algorithm. It trains the classifier to predict the class of the new data. SVM is invented by Vapnik. Due to its remarkable features it has many applications like face detection, classification of images, handwriting recognition and many more. It main aim to focus on finding optimal training datasets. It assures high accuracy even if the dataset is small [16]. SVM is also used for web attacks like sql injections, cross-site scripting, etc.

### F. Logistic Regression

It is a classification algorithm under supervised learning. It predicts output in binary forms (1 == Yes/True, 0 == No/False) in given dataset for input elements. There are numerous application of logistic regression such as geographical image processing, financial forecasting, image segmentation and categorization and many more. Dummy variables are used to represents outputs. It can be ordinal, binomial and multinomial. Ordinal works on dependent variables that are ordered. Binomial deals with results which have two outcomes (0/1, Yes/No, True/False). Last, multinomial deals with results which have more than two outcomes; for examples ("disease A" vs "disease B" vs "disease C").

TABLE I. RELATED WORK

| Method | Description | Reference |
|---|---|---|
| Bayesian classification technique | based on the uncertain data. They take 20 data sets from UCI repository and apply uncertain Bayesian classification and prediction technique | Ren, J., Lee, S. D., Chen, X., Kao, B., Cheng, R., & Cheung, D. (2009, December) [5] |
| K nearest neighbor | query dependent ranking. They first consider the online method and next consider two offline methods which create a ranking model to enhance the efficiency of ranking in advance and approximation are accurate in terms of difference in loss of prediction | XiuboGeng et al [6] |
| SVM and KNN algorithm | technique that produces the fault detection of engine journal bearing. | A.Moosavianet at [8] |
| SVM algorithm | minimizing the misclassification; they employ a risk decision rule of empirical risk Minimization (ERM) for a non-separable sample in MATLAB. Shows computational result is better than the SVM | Xuemei Zhang et al [7] |
| Logistic regression, Back Propagation Neural network, support vector machine algorithms | Proposed a novel method for medical problem. They combine the PSO and C4.5, where PSO is used in the feature selection technique and C4.5 adopts PSO fitness function for classification by using five datasets from UCI repository. | Mena ChanaTasi et al [9] |
| Support vector (SVM) algorithm | Based on medical diagnosis. they use new methodology in which training set is divided into two subsets, first subset is used to train SVM with RBF kernel, other subset is used to train other SVM with polynomial kernel | Savvaskaratsiolis et al [10] |

## IV. Comparative Study

This comparative study shows (figure 2) that the survey of machine learning algorithms by IBM. Figure 3 consist the most popular language for data mining.

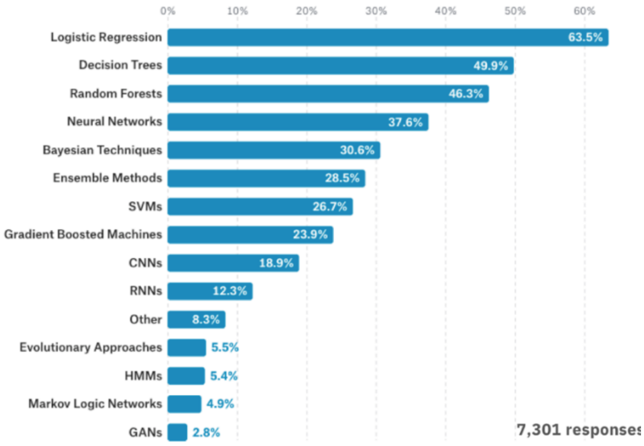

**Figure 2.** Popular algorithms survey by IBM

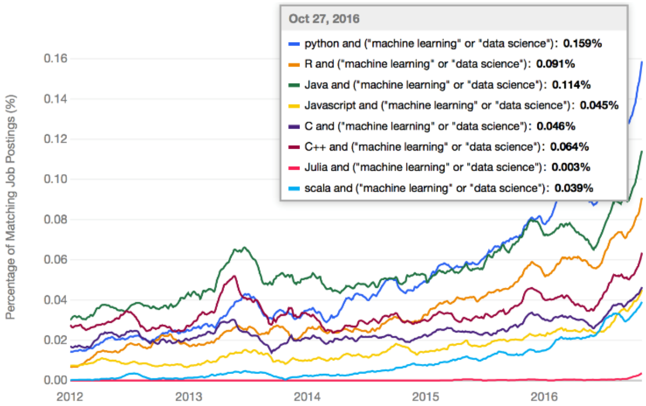

**Figure 3.** Trend of popular language used for data analysis

## V. CONCLUSION

This paper specifies various pythons libraries and classification techniques used in many fields, such as Decision Tree, K-means clustering, Random forest, Logistic regression, neural network etc. Generally, Decision trees and Support vector machines have different functionality, where one is predicted and other is not or vice versa. On the other hand, decision trees and rule classifiers have a similar operational profile. Various algorithms will be combined for classifying the data set. This paper provides compressive overview of various classification techniques used in different fields of data mining. In any field one classification technique is more useful than another. This paper presents various classification techniques. One of the above techniques can be selected based on the required application conditions.

## VI. REFERENCES

[1] Andreas C. Müller & Sarah Guido. 2016. Introduction to Machine Learning with Python. O'Reilly Media.

[2] Gavin Hackeling. 2014. Mastering Machine Learning with scikit-learn. Packt Publishing.

[3] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort. (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 2825-2830.

[4] Eli Bressert. 2012. SciPy and NumPy. O'Reilly Media.

[5] Ren, J., Lee, S. D., Chen, X., Kao, B., Cheng, R., & Cheung, D. (2009, December). Naive bayes classification of uncertain data. In Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on (pp. 944-949). IEEE.

[6] Geng, X., Liu, T. Y., Qin, T., Arnold, A., Li, H., & Shum, H. Y. (2008, July). Query dependent ranking using k-nearest neighbor. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 115-122). ACM

[7] Zhang, X., & Yang, L. (2012). Improving SVM through a risk decision rule running on MATLAB. Journal of Software, 7(10), 2252-2257

[8] Moosavian, A., Ahmadi, H., & Tabatabaeefar, A. (2012). Journal-bearing fault detection based

on vibration analysis using feature selection and classification techniques.

[9] Taniar, D., & Rahayu, W. (2013). A taxonomy for nearest neighbour queries in spatial databases. Journal of Computer and System Sciences.

[10] Karatsiolis, S., & Schizas, C. N. (2012, November). Region based Support Vector Machine algorithm for medical diagnosis on Pima Indian Diabetes dataset. In Bioinformatics & Bioengineering (BIBE), 2012 IEEE 12th International Conference on (pp. 139-144). IEEE

[11] Eesha Goel, Er. Abhilasha (2017, January). Random Forest: A Review, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X DOI: 10.23956/ijarcsse/V7I1/01113

[12] Seema Sharma, Jitendra Agrawal , Shikha Agarwal , Sanjeev Sharma (2013). Machine Learning Techniques for Data Mining: A Survey, 2013 IEEE International Conference on Computational Intelligence and Computing Research, 978-1-4799-1597-2/13

[13] Shi Na, Guan yong, Liu Xumin (2010). Research on k-means Clustering algorithm: An Improved k-means Clustering Algorithm, Third International Symposium on Intelligent Information Technology and Security Informatics IEEE, ISSN: 978-0-7695-4020-7/10 DOI 10.1109/IITSI.2010.74

[14] Fahim A M,Salem A M,Torkey F A, “An efficient enhanced k-means clustering algorithm” Journal of Zhejiang University Science A, Vol.10, pp:1626-1633,July 2006.

[15] Ms. Sonali. B. Maind, Ms. Priyanka Wankar (2014, January), Research Paper on Basic of Artificial Neural Network. International Journal on Recent and Innovation Trends in Computing and Communication. ISSN: 2321-8169

[16] Abhishek Gupta, Ankit Jain, Samartha Yadav. (2018). Literature survey on detection of web attacks using machine learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2456-3307.