Python Libraries and Packages for Data Mining-A Survey

S.Sangeetha[1], Dr. S. Saradhambekai[2]
[1]*PG Student, Department of Information Technology, PSG College of Technology, Coimbatore-4, India*
[2]*Assistant Professor (Sr. Gr), Department of Information Technology, PSG College of Technology, Coimbatore-4, India*

***Abstract*- Python is the one of the scripting language that is simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. It is also an interpreter, object oriented and high level programming language with dynamic sematics. Using python packages for data mining provide secure, customer acquisition, and improvement in planning acquisition. It helps for analyser to analyse the data for particular organisation.**

In this paper, the survey of various papers that perform with python modules and libraries for data mining are attached and analyzed with metrics like performance, reliability and stability because of using python packages and libraries.

Index Terms- scripting, platform independent, flexible, troubleshooted, performance, reliability, stability, secure, ease of use.

I. INTRODUCTION

This survey paper depiction the various python packages that are been used in data mining which drastically increases the performance of mining objects.

II. SIGNIFICANCE OF PYTHON

The significance of python is described as follows,

a. Python is a portable language
b. Python is a Beginner language.
c. Python is an Object-oriented scripting language
d. Python is a portable language
e. Python is high-level programming.
f. Python provides interfaces to all databases.
g. Python is an interactive language.
h. Python supports GUI Programming language.
i. Python supports very portable and cross-platform compatible on UNIX, Windows and Macintosh.

III. SIGNIFICANCE OF DATA MINING

The significance of networking is described as follows,

a. Increase decision making.
b. Improves security risk posture.
c. Improves forecasting and planning.
d. Competitive advantage.
e. Customer acquisition.
f. Cost reduction.
g. Expand customer relationship.
h. New revenue streams.
i. Development of new products

IV. PYTHON LIBRARIES AND PACKAGES FOR DATA MINING

a. NumPy

NumPy isa basic packages in python for scientific computing. NumPy provides an extension to the Python programming language for adding large, multi-dimensional arrays and matrices, along with a large library of high-level mathematical functions to operate on the given arrays.

b. SciPy

SciPy is free and open-source software for engineering, mathematics, and science. The SciPy library based on NumPy, which provides convenient and fast N-dimensional array manipulation. The SciPy library is work with NumPy to build arrays, and makes user-friendly and resourceful arithmetical routines such as routines for mathematical combination and optimization. Mutually, they scamper on all admired operating systems, are quick to set up, and no cost of charge.

c. Pandas

Pandas is a fast, elastic, and communicative data structures consider to make running with “relational” or “labeled” data both trouble-free and sensitive in python package. It aims to be the deep-seated high-level building block for doing realistic in actual

world data investigation in Python. Furthermore, it has the broader target of suitable to all nearly powerful and flexible open source data analysis / operation tool available in any language. It is already well on its way toward this purpose.

d. Matplotlib

Matplotlib is a scheming documents for the Python training language and its NumPy algebraic mathematics expansion. It provides an API which is object oriented for embedding plots into other applications using widespread-purpose GUI toolkits like Qt, or GTK+, wxPython,. There is also a technical "pylab" crossing point based on a state machine (like OpenGL), planned to directly look like that of MATLAB. SciPy makes use of matplotlib.

e. IPython

IPython is a power shell for interactive computing in several programming languages, originally developed for the Python programming language, that offers better introspection, rich medium, extra shell syntax, tab conclusion, and rich past.

f. SCIKIT-Learn

The scikit-learn project in progress as scikits.learn, a Google Summer of Code project by David Cournapeau. Its name stems from the concept that it is a "SciKit" (SciPy Toolkit), a separately-developed and circulated third-party expansion to SciPy. The starting codebase was well along widely modified by former designer. Of a range of scikits, scikit-learn as well as scikit image were described as "well maintained and well liked".

V. RELATED RESEARCH WORKS

A.EARLY HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

[1] focused on using different algorithms and mixture of some target attributes for efficient heart attack prediction using data mining.

This is built to run on data mining weka 3.6.6 .

It has three methods of operation namely,

a. Decision Tree: It is a simple and easy good classifier for processing. No need of domain knowledge but have high dimensional data to handle and produce result easily.

b. Naïve Bayes: This produce a statistical classifier with attributes but no dependency required. The main advantage is work without any Bayesian method.

c. Neural Networks: It is a computational model depends on biological neural network. The feed-forward neural networks the neurons of the first layer forward their output to the neurons of the second layer.

NumPy is used for build to run datas on statistical analysis on weka used to extract hidden patterns from large datasets for predicting the cause for heart attack.

B. Survey on data mining techniques for disease prediction

[2] predicting the heart attack which causes to death condition. In these paper the classification, prediction, clustering, ensemeble learning and boosting.

It used seven approaches for prediction namely,

a. Decision Tree(J48): It predict the class and target value based on decision rules. Following same representation of SOP for every classes. It handles the missing value during prediction.J48 supports tree pruning.

b. Support Vector Machine: supervised knowledge classifier characterize by hyper plane.

The hyper plane are

- Non Linear SVM classifier
- Linear SVM classifier

These are utilized in z-score standardized in numeric form.

c. Naïve Bayes: It records the path which is specified in the class. The highest probability class is consider as most possible class.

d. Random Forest: It is used for both classification and regression purpose. In classification issues, the group of trees chooses in favor of the most outstanding class. In the regression problem, their responses are averaged to obtain an estimate of the reliant variable.

e. Adaboost: It's basic principle is to in shape train weak learners. The prediction from one of classes are then joined through a weighted greater part of vote (or sum) to bring the last prediction.

f. MultiLayer Percepteron: Make use of various layers of the neural network is formed by with the position of different parameters which are chosen to

adjust the models with the assist of correlation between parameters and prediction of the disease

g. Logistic Regression: It measures the connection between dependent and independent values or variables in logistic function.

In these paper matplotlib library is used for predict diseases was briefly explained.

C.Heart Disease Prediction using Data Mining Classification

[3] classifying heart diseases based on classification and evaluate based on particular classifier to identify the accuracy .

Finding accuracy with four application Explorer, Experimenter, Knowledge Flow and Simple CLI.

a. Naive Bayes

Naive Bayesian model is trouble-free to build, with no difficult iterative parameter assessment which makes it particularly useful for very large datasets.

b. Artificial Neural Network

With the help of artificial neural network, a set of inputs are mapped into set of proper output. The nonlinear activation functions uses three or more layers of neurons which is more powerful than the perception in that it can decide data that is separable by a hyper-plane.

It uses python libraries such as inbuilt and external modules to analyse perfect accuracy.

D. Multi Disease Prediction using Data Mining Technique

[4] For predicting disease number of test should be required from the patients details.

a. Naïve Bayes: It has an independence assumptions between predictors. It is easy to build,useful for very large datasets

b. Decision Tree: Decision tree is a classifier in the structure of a tree structure where each node is either a leaf node, indicating the value of the target attribute or class. Observe data and its rules that will be used to make predictions.

E: Prediction of Cardiac Disease Based on Patient's Symptoms

[5] Quick detection of all possible symptoms and signs which cause cardiac diseases detect using previous data set of patients as well as current data received from user at actual time.

The collected data are stored in a pre-processing dataset by outlier analysis.

The judgement based on the validation on the methods

AUC: It is an area under the curve. It is used for classification analysis in order to decide which of the used models predicts the classes best. An example of its application are ROC curves.

CA: Classification Accuracy is a method to know the accuracy of the classifier on valid testing data. It is the ratio of correct predictions made to total predictions made.

Recall: The training data helps to classify the model which was used again to test the recalling abilities of the modelled classifier.

These techniques are implement in the Python language with Sci-Kit learn to analyse based on the code with datasets.

Summary of the Research Related Works – Table 1

S. no	Title	Author Name	Description	Python packages used	Merits	Demerits
A	Early Heart Disease Prediction Using Data Mining Techniques	Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar	Predicts the patient having heart disease using MAFIA algorithm	NumPy package is used.	Keeps minimum space for size and provide fast.	Not scalable for large datasets.
B	Survey on data mining techniques for disease prediction	Durga Kinge, S. K. Gaikwad	Analysing the different diseases using various algorithm in data mining	Matplotlib is used for analysed the datasets and classify the classes in python	Programmig interface has Matlab way, build in code with default plot	It strongly depends on other packages and work for python only.
C	Review of Medical Disease Symptoms Prediction Using Data Mining Technique	Rahul Deo Sah, Dr. Jitendra Sheetalani	Techniques that can be used for heart diseases classification and find the accuracy of selected classifier algorithm	Uses SciPy python modules for Explorer, Experimenter and Knowledge flow	Random numbers can generate. t uses multidimensional container datas and supports many operating system.	It strongly depends on other packages library packages and work for python only.
D	Multi Disease Prediction using Data Mining Technique	K.Gomathi, Dr. D. Shanmuga Priyaa	Identifying the number of test to find the disease by using patient	IPython libraries packages for testing the classification	Faster acessing, plot inline on graph, supports text and allow other media accessing	It has a combination of other package library.
E	Prediction of Cardiac Disease Based on Patient's Symptoms	Prabakaran.N and Kannadasan.R	Identify the disease based on the previous dataset and current dataset of the user.	Scikit Learn package is used.	Hygenic API, tough, fast, easy accessing, supported, released under a permissive license	It does not gives importance to statistics

VI. CONCLUSION

This survey paper covers few python libraries and modules that are been used in data mining. With the help of python, it reduce many issues such as lack of performance, lack of trustworthiness. The code size has been relatively reduced such that it enhances the throughput and reduces the memory usage for the data.

Importing the various python data mining libraries have made a programmed and elastic environment for mining analyser to work.

Since python has inbuilt libraries and also provides a provision to include external modules, it is easier to code the project.

Using the python libraries, it is easy to detect and analyse the disturbance in the data and it is easy to identify and remove the disturbance from the data [1][2][3].

In these datasets can be used as a pre-processing of data .It can be easily modified the data for the current dataset[5]

Datasets are tested and classify based on the iteration and analysed the data on the given performance of the given data.

Thus, python provides various libraries and modules that can be used in data mining which helps the data analyser to easily code, deploy and monitor the network from intrusions and building an efficient data mining platform with python [4].

REFERENCES

[1] Methaila, A., Kansal, P., Arya, H., & Kumar, P. (2014). Early heart disease prediction using data mining techniques. Computer Science & Information Technology Journal, 53-59.

[2] Durga Kinge, S. K. Gaikwad. Survey of data mining techniques on disease prediction. International Research Journal of Engineering and Technology, Vol.5 Issue.1, Jan- 2018

[3] Rahul Deo Sah, Dr.Jitendra Sheetanalai. Review of Medical Symptoms Prediction Using Data Mining Technique. IOSR Journal of Computer Engineering (IOSR-JCE), Volume 19, Issue 3, ver.1 (May-June.2017),59-70.

[4] Gomathi K., Dr D. Shanmuga Priyaa. Multi Disease Prediction using Data Mining Technique International Journal of System and software Engineering. Volume 4 Issue 2.December 2016

[5] Prabakaran, N., & Kannadasan, R. (2018, April). Prediction of Cardiac Disease Based on Patient's Symptoms. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 794-799). IEEE.