Python Libraries and Packages for Machine Learning – A Survey

M.Nivethitha[1], Dr. S. Sarathambekai[2]
[1]*PG Student, Department of Information Technology, PSG College of Technology, Coimbatore-4, India*
[2]*Assistant Professor (Sl. Gr), Department of Information Technology, PSG College of Technology, Coimbatore-4, India*

***Abstract*- Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is a programming language that work quickly and integrate systems more efficiently.**

In this paper, the survey of different papers that used python modules and libraries for machine learning are taken and analysed with metrics like average accuracy, performance, reliability and stability because of using python packages and libraries.

Index Terms- python, accuracy, reliability, efficiency, ease of use.

I. INTRODUCTION

In this survey paper the different python packages that are been used in machine learning algorithms is depicted and along with their performance metrics and average accuracy achieved.

II. CHARACTERISTIC FEATURES OF PYTHON

The characteristic features of python are as follows:

a. Interactive
b. Interpreted
c. Modular
d. Dynamic
e. Object-oriented
f. Portable
g. High level
h. Extensible in C++ & C

III. ADVANTAGES OF MACHINE LEARNING

a. Easily identifies trends and patterns
b. No human intervention needed (automation)
c. Continuous Improvement
d. Handling multi-dimensional and multi-variety data
e. Wide Applications

IV. PYTHON LIBRARIES AND PACKAGES FOR MACHINE LEARNING

Python libraries that are used in Machine Learning are:

a. Numpy

NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning.

b. Scipy

SciPy is a very popular library in Machine Learning as it contains different modules for optimization, linear algebra, integration and statistics. The SciPy is one of the core packages that is very useful for image manipulation.

c. Scikit-learn

Skikit-learn is one of the most popular ML libraries for classical ML algorithms. It is built on top of two basic Python libraries, viz., NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms.

d. Theano

Theano is a popular python library that is used to define, evaluate and optimize mathematical expressions involving multi-dimensional arrays in an efficient manner. It is extensively used for unit-

testing and self-verification to detect and diagnose different types of errors. Theano is a very powerful library that has been used in large-scale computationally intensive scientific projects for a long time.

e. TensorFlow

TensorFlow is a very popular open-source library for high performance numerical computation developed by the Google Brain team in Google. As the name suggests, Tensorflow is a framework that involves defining and running computations involving tensors. It can train and run deep neural networks that can be used to develop several AI applications. TensorFlow is widely used in the field of deep learning research and application.

f. Keras

It is a high-level neural networks API capable of running on top of TensorFlow, CNTK, or Theano. Keras allows for easy and fast prototyping.

g. PyTorch

PyTorch is a popular open-source Machine Learning library. It has an extensive choice of tools and libraries that supports on Computer Vision, Natural Language Processing (NLP) and many more ML programs.

V. RELATED RESEARCH WORKS

A. Enabling Adaptability in Web Forms Based on User Characteristics Detection Through A/B Testing and Machine Learning

[1] has stated a solution for improving users' performance in completing large questionnaires through adaptability in web forms.

The goal of this paper is to present a new approach for enabling adaptability in web-based systems using A/B testing methods and user-tracking and machine-learning algorithms that could lead to improving user performance in completing a (large) web form, validating the obtained results through statistical tests. The python package NumPy is used here.

Predictive models and clustering were used for data processing.

a. The first attempt was based on using all the data together focusing in primary variables (excluding those that have too many void values).
b. The second one was based on using all variables and derived variables (constructed from primary ones).
c. The third attempt was based on creating separated predictive models depending on the vertical.

Hence this paper proposed that adaptability can be achieved by detecting users behaviours, preferences, and profiles using machine-learning techniques and offering the best user interface and user experience to each kind of user detected.

B. Learning phrase representations using rnn encoder–decoder for statistical machine translation

[2] proposed that the model has a semantically and syntactically meaningful representation of linguistic phrases.

The proposed RNN Encoder–Decoder with a novel hidden unit is empirically eval-uated on the task of translating from English to French.

SciPy package is used for integration and optimisation of various languages.

The two methods that is used here are as follows

Recurrent Neural Networks:

A recurrent neural network (RNN) is a neural network that consists of a hidden state and an optional output which operates on a variable length sequence.

Statistical Machine Translation:

The system (decoder, specifically) is to find a translation in a given source sentence, which maximizes translation model and the latter language model.

This paper proposed a solution for mapping from a sequence of an arbitrary length to another sequence, possibly from a different set, of an arbitrary length.

C. A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning

[3] presents the imbalanced-learn API, a python toolbox to tackle the curse of imbalanced datasets in machine learning.

The python package scikit-learn is used as a major support for the toolbox used.

The two methods used here are as follows

Sampling:

The process of reducing the number of samples is called sampling.

The implemented methods can be categorized into 2 groups:
(i) fixed under-sampling and
(ii) cleaning under-sampling
Ensemble-learning:
Ensemble methods offer an alternative to use most of the samples.
This paper presented the foundations of the imbalanced-learn toolbox vision
and API.

D. Deep Learning on GPUs with Python
[4] states that a symbolic manipulation engine geared towards optimizing and executing expression graphs on tensors.
A Graphical Processing Unit (GPU) device (when present) to evaluate such mathematical expressions as quickly and accurately as possible.
Theano is a Python library improve both execution time and development time of machine learning applications, especially deep learning algorithms.
There are four steps to using Theano in a Python program:

1. Declare symbolic input variables.
2. Construct a symbolic expression graph.
3. Compile one or more functions to evaluate particular expressions.
4. Call those functions to evaluate expressions for particular input values.

Thus this paper emphasis on internal computations performed on a GPU. The speed of Theano's compiled functions gradient descent in DBN and convolutional architectures across several implementations.

E. Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python
[5] states the challenging format of the tweets which makes the processing difficult in the twitter sentiment analysis.
Natural Language toolkit (NLTK) is a library in python, which provides the base for text processing and classification. Operations such as tokenization, tagging, filtering, text manipulation can be performed with the use of NLTK.
The two methods that are used here are as follows
Pre-processing in Python
This process performs converting and removing of the unwanted data in the data set.
Feature Extraction
Term frequency-Inverse Document frequency is an efficient approach. TF-IDF is a numerical statistic that reflects the value of a word for the whole document.
Hence this paper focuses on analysing the sentiments of the tweets and feeding the data to a machine learning model in order to train it and then check its accuracy.

SUMMARY OF THE RESEARCH RELATED WORKS – TABLE 1

S.NO	TITLE	AUTHOR NAME	DESCRIPTION	PYTHON PACKAGES USED	MERITS	DEMERITS
A.	Enabling Adaptability in Web Forms Based on User Characteristics Detection Through A/B Testing and Machine Learning	JUAN CRUZ-BENITO et al.,	Improving users performance in completing large questionnaires through adaptability in web forms	NumPy	White-box approaches, performance, promising results, singular changes	Size of the device and browser window.
B.	Learning phrase representations using rnn encoder–decoder for statistical machine translation	Kyunghyun Cho et al.,	semantically and syntactically meaningful representation of linguistic phrases	SciPy	improves the translation performance, continuous space representati	Different language support
C.	A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning	Guillaume Lemaitre et al.,	A python toolbox to tackle the curse of imbalanced datasets in machine learning	imbalanced-learn API, Sci-kit learn	Quality assurance, Continuous integration, Community-based development, Documentation, Project relevance	prototype/instance selection, generation, and reduction
D.	Deep Learning on GPUs with Python	James Bergstra et al.,	symbolic manipulation engine geared towards optimizing and executing expression graphs on tensors	Theano	improve both execution time and development time, best strategy for computing	highly-documented
E.	Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python	Bhumika Gupta et al.,	challenging format of the tweets which makes the processing difficult in the twitter sentiment analysis	Natural Language toolkit (NLTK), NumPy, SciPy	Limited tweet size, Special character and digit removal, Spelling correction, Expansion of slangs and abbreviations	lacks the dimension of diversity in the data

CONCLUSION

This survey paper comprises of various python libraries and modules that are been used in machine learning.

Importing the various python machine learning libraries have made the automated and flexible environment for machine learning algorithms.

Python is very popular for its code readability and compact line of codes. It uses white space inundation to delimit blocks.

It is favoured for complex projects, because of its simplicity, diverse range of features and its dynamic nature.

Thus, python provides various libraries and modules that can be used in machine learning to easily code, deploy applications efficiently using machine learning algorithms.

REFERENCES

[1] Cruz-Benito, J., Vázquez-Ingelmo, A., Sánchez-Prieto, J. C., Therón, R., García-Peñalvo, F. J., & Martín-González, M. (2018). Enabling adaptability in web forms based on user characteristics detection through A/B testing and machine learning. IEEE Access, 6, 2251-2265.

[2] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

[3] Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research, 18(1), 559-563.

[4] Bergstra, J., Bastien, F., Breuleux, O., Lamblin, P., Pascanu, R., Delalleau, O. & Bengio, Y. (2011). Theano: Deep learning on gpus with python. In NIPS 2011, BigLearning Workshop, Granada, Spain (Vol. 3, pp. 1-48). Microtome Publishing.

[5] Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., & Badhani, P. (2017). Study of Twitter sentiment analysis using machine learning algorithms on Python. International Journal of Computer Applications, 165(9), 0975-8887.