**ML1819 Research Assignment 1**

Team 02

Task 101

Rajesh Burla - 18306485

Akashdeep Singh Lamba - 18305063

Shanmukha Sai Ram Pavan - 18305688

Each student picked up one library and implemented the 3 selected algorithms end-to-end in the assigned library, while constantly communicating progress over a Slack channel. We each maintained shared git repos to share code and collectively work. All three of us then compared our results and then discussed the contents of this document. We realized we were supposed to work in a single repo so we migrated our code to a shared repo in a group with write access to only the three of us. This also means the contributor graph isn’t a perfect reflection of our activity.

Word count: ###

<https://gitlab.scss.tcd.ie/ML1819-team-02/ML1819--task-101--team-02>

<https://gitlab.scss.tcd.ie/ML1819-team-02/ML1819--task-101--team-02/graphs/master>

###[Screenshot]

Survey of popular Machine Learning Libraries

A comparative analysis of Tensorflow, Sklearn and Weka by implementation of standard algorithms

Rajesh Burla  
MSc Computer Science - Intelligent Systems (2018-19)  
 Trinity College Dublin  
 Dublin, Ireland  
[burlar@tcd.ie](mailto:burlar@tcd.ie)

Akashdeep Singh Lamba  
MSc Computer Science - Intelligent Systems (2018-19)  
 Trinity College Dublin  
 Dublin, Ireland  
 [lambaa@tcd.ie](mailto:lambaa@tcd.ie)

Shanmukha Sai Ram Pavan   
MSc Computer Science - Intelligent Systems (2018-19)  
Trinity College Dublin  
 Dublin, Ireland  
 [parvarths@tcd.ie](mailto:parvarths@tcd.ie)

ABSTRACT

This paper briefly compares 3 of the popular machine learning libraries by analysing implementation of linear regression, support vector machine classification and k-nearest neighbour algorithms.

1 INTRODUCTION

Due to growing interest in machine learning, several open-source libraries have become popular for different reasons, each preferred over the other for either ease-of-use, speed, accuracy or other features like GPU support. It is worth asking whether all libraries perform similarly under approximately same conditions or not, and this paper addresses this question.

2 RELATED WORK

Analysing the experiments of Bhuvan M Shashidhara et al. in [1] the results shows that Scikit-Learn is best fit for data in comparison with Weka and Apache Spark frameworks.

Scrutinizing the results of the Google Brain team [2] reveals that TensorFlow is a flexible dataflow representation that enables power users to achieve excellent performance and scalability.

3 METHODOLOGY

In this experiment, three different Machine learning algorithms and frameworks were chosen.

The dataset used was Google Play Store Apps: web scraped data of 10k Play Store apps for analysing the Android market. This dataset contains all the details of the applications on Google Play. There are 13 features that describe an individual app. [3]

3.1 Data Pre-processing

The dataset in the original form obtained from Kaggle wasn’t fit for direct use. In order to make it suitable for the experiment, sparse columns were eliminated, string columns Category and Genres encoded as numbers, and a new column called Rated 4.4 or more derived from Rating was created for use with SVM. Further, scaling was left to implementation specific code.

3.2 Deployment on different platforms

**Weka**: It contains robust sequential implementations of many machine learning algorithms, and ships with an easy to use GUI.

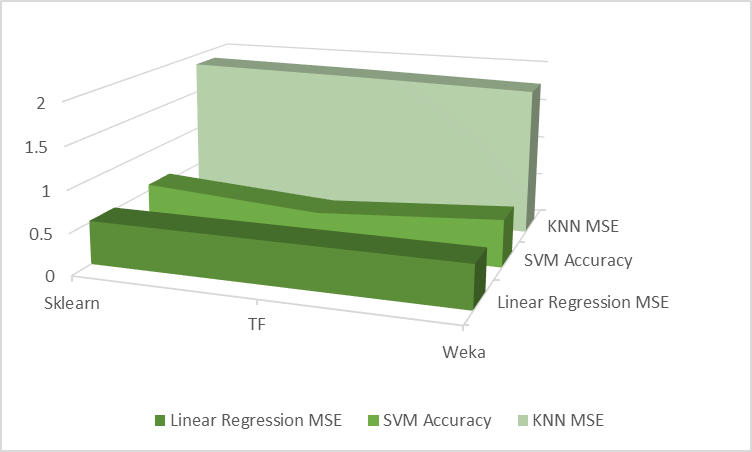
**Scikit-Learn**: Due to it’s in-memory processing, it is very fast for smaller data which can be easily loaded into memory.

**TensorFlow**: TensorFlow receives data in the form of Tensors, which are data in the form of arrays of dimensions and ranks. It supports distributed execution over GPUs and CPUs.

1.3 Machine learning Algorithms

Three machine learning algorithms are chosen, Linear Regression, Support Vector Machines and KNN to test Root Mean Square Error and accuracy with the trained dataset on various frameworks.

4 RESULTS & DISCUSSION



5 LIMITATIONS & OUTLOOK

While the experiment conducted within the prevailing constraints provides an overview of comparison between the three libraries, it is not by any means a comprehensive analysis. Metrics such as training speed haven’t been considered, and more sophisticated algorithms haven’t been tested.

ACKNOWLEDGMENTS

This analysis was conducted as part of the 2018/19 Machine Learning module CS7CS4/CS4404 at Trinity College Dublin [4].

BIBLIOGRAPHY

[4] Joeran Beel and Douglas Leith. Machine Learning (CS7CS4/CS4404). Trinity College Dublin, School of Computer Science and Statistics. 2018.

REFERENCES

[1] Bhuvan M Shashidhara et al. “Evaluation of Machine Learning Frameworks on Bank Marketing and Higgs Datasets” IEEE 2015 DOI: 10.1109/ICACCE.2015.31

[2] The team Google Brain’s “TensorFlow: A system for large-scale machine learning” 12th USENIX Symposium on Operating Systems Design and Implementation in 2016.

[3] Google Play Store Apps dataset <https://www.kaggle.com/lava18/google-play-store-apps>