**Analysis of Machine Learning Algorithms**

Yugandhar G, Prashanth M, Praveen M, Vinyas R

Mrs Sudhamani M J

Asst. Prof, Guide

Department of Computer Science & Engineering

***Abstract: A comprehensive study is done on Finger vein, Fisher iris, Agar Wood and Glass datasets with SVM, k-NN, NN, k-Means, Random Forest, Linear regression, Logistic regression and DBSCAN algorithms. The performance of different classification and clustering algorithms are discussed aiming at identifying the most appropriate features. The accuracy for each of the algorithm was calculated and analyzed. As a result we identified most accurate algorithm which***

***predicts data accurately with compared to other algorithms.***

**Keywords:** SVM, KNN, DBSCAN, Regression, Testing, Training, and Dataset.

**I.** INTRODUCTION

Machine learning is a branch of science that deals with programming the systems in such a way that they automatically learn and improve with experience. Machine learning programs are also designed to learn and constantly improve themselves over time based on the new data the encounter, its neural network learning power has been shown in different prediction methods. Machine learning can be divided into supervised, semi supervised and unsupervised learning. Supervised learning is that when the computer is first made to learn a set of patterns and then recognize the patterns that may be similar to that on a training set. Then using the same knowledge the computer identifies patterns on any new data set. Unsupervised learning does not involve “teaching” the computer. It is intelligent enough to automatically discover any patterns. Regression is one of the sub-types of semi-supervised learning. It involves establishing relationships among the various variables which is improved using a measure of error in the predictions made by the system.

*Classification* is a machine learning technique that uses known data to determine how the new data should be classified into a set of existing categories.

*Clustering* is used to form groups or clusters of similar data based on common characteristics.

*Regression* is a parametric technique used to predict continuous variable given a set of independent variables.

II. PROBLEM STATEMENT

Large scale software systems challenge almost every activity in the software development life-cycle, including tasks related to eliciting, analyzing, and specifying requirements. Project revolves around comprehensive and comparative study of several Machine learning algorithms involving real time data sets. This template a streamlined process that focuses on learning about the problem, a good solution, and doing so very quickly.

**III**. SYSTEM DESIGN

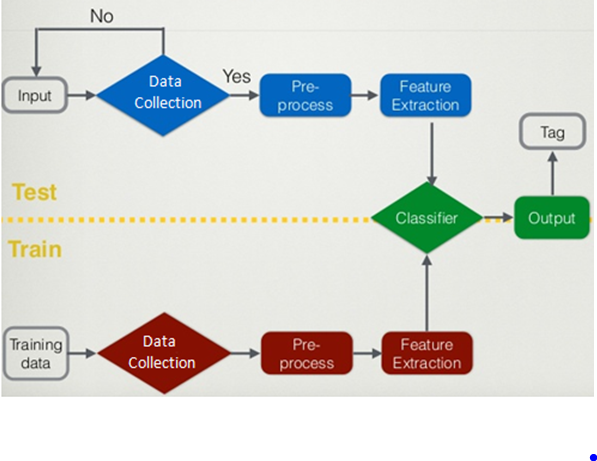
Input: Raw data from in stream data lines or external files are supplied to the algorithm

Data Collection: It is the process of gathering and measuring information on targeted variables in an established systematic fashion, which then enables one to answer relevant questions and evaluate outcomes.

Pre-processing: It refers to the transformations applied to convert the raw data into a clean data set before feeding it to the algorithm.

Feature Extraction: It is a type of dimensionality reduction that efficiently represents interesting parts of an image as a compact feature vector. This approach is useful when image sizes are large and a reduced feature representation is required to quickly complete tasks such as image matching and retrieval.

Classifier: It is the process of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.



VI. SYSTEM REQUIREMENTS

1. Hardware requirements

* Processor: Any Intel or AMD x86-64 processor.
* RAM: 4 GB or better.
* Hard Drive: 22 GB or better.

1. Software requirements

* Operating System: Windows7.
* MATLAB R2017a
* Pycharm Version: 2018.1.2

**V**. IMPLEMENTATION

**Support Vector Machine**: It is a classification method. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

**K nearest neighbours**: It is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbours.

## Linear Regression: It is used to estimate real values based on continuous variable.

## DBSCAN: It is a [density-based clustering](https://en.wikipedia.org/wiki/Cluster_analysis#Density-based_clustering) algorithm given a set of points in some space, it groups together points that are closely packed together, marking as outliers points that lie alone in low-density regions.

## Logistic Regression: It is used to estimate discrete values based on given set of independent variable.

## Random forest: It is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees.

## K-Means: It is a type of unsupervised algorithm which solves the clustering problem. Its procedure follows a simple and easy  way to classify a given data set through a certain number of  clusters

VII. TEST ANALYSIS

The performance of the classifiers can be compared according to certain metrics like accuracy, specificity, sensitivity, training time etc. A confusion matrix forms the basis from which different parameters can be calculated. The number of instances accurately or inaccurately predicted by a classification model can be tabulated in the form of a confusion matrix. The confusion matrix is generally represented by 4 values which are TP, FN, FP and TN as shown in Table I. The parameters are discussed in brief below.

*True positive (TP) —* It indicates the instances which are predicted as normal correctly.

*False negative (FN) —* It indicates wrong prediction i.e. it detects instances which are attacks in reality, as normal.

*False positive (FP) —* It gives a hint of the number of detected attacks which are normal in reality.

**VIII.** CONCLUSION

Algorithms implemented in this project provides a basis for machine learning, whose test cases and outcomes can be used for further research and optimizing the algorithms. We have compared the performance of various classification and clustering algorithms with different datasets. Experiments were carried out using different real time datasets and UCI data sets to analyze and compare their classification performance. Feature extraction concept used here can be implemented in the field of image processing which will result in adaptable and resilient image processing technology. A multimodal scheme where several approach systems will be combined to form a unit identification system can be studied. As a future work, the proposed approach will be tested on other datasets. Focus can be given to integration and comparison between the ROC (Receiver Operating Characteristics ) curves. The proposed approach contains several phases, namely, acquisition, enhancement, feature extraction and classification.

**IX.** REFERENCES

1. P. Kang Ryoung “ finger vein recognition by combining global and local features based on SVM,” Computing and Informatics, Vol. 30,295–309,2011.
2. R. Agrawal. “K-Nearest Neighborn for Uncertain Data”. International Journal of Computer Applications (0975-8887). 2014. Vol. 105 No. 11

pp 13-16.

1. Singh, Aruma, and Patel, Smita, Shukla,. “Applying Modified KNearest Neighborn to Detect Threat in Collaborative Information Systems”.
2. Itskovich, Lev, and Sergei Kuznetsov. "Machine learning methods in character recognition." In *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, pp. 322-329. Springer Berlin Heidelberg, 2011.
3. Friedman, J.H.: Stochastic gradient boosting. Computational Statistics and Data Analysis 38(4), 367–378 (2002).
4. C. L. P. Chen and C.-Y. Zhang, “Data-intensive applications, challenges,

techniques and technologies: A survey on big data,” Information Sciences,

vol. 275, pp. 314 – 347, 2014.

1. J. Tang, C. Deng, and G. B. Huang, “Extreme learning machine for

multilayer perceptron,” IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 4, pp. 809–821, April 2016.

1. M. Chen, Z. E. Xu, K. Q. Weinberger, and F. Sha, “Marginalized denoising

autoencoders for domain adaptation,” CoRR, vol. abs/1206.4683,

2012.

1. Korb, Kevin B., and Ann E. Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.