Contents

Chapter 1

Introduction

The aim of this report is to use the machine learning techiques (Logistic Regression ,KNN and SVM) on 'shuttle' dataset[1], using scikit library of python[2] and compare between different machine learning techniques in terms of accuracy , traning time and testing time. Results are obtained through benchmark which is designed for this purpose.

Chapter 2

Statistics

2.1 Description

According to the description of the 'shuttle' dataset [1] contains 9 attributes all of which are numerical. The first one being time. The last column is the class which has been coded as follows:

1. Rad Flow
2. Fpv Close
3. Fpv Open
4. High
5. Bypass
6. Bpv Close
7. Bpv Open

2.2 Histogram

This Histogram of ’Shuttle’ dataset[1] for the test sample according to the above classes shows the following classification:

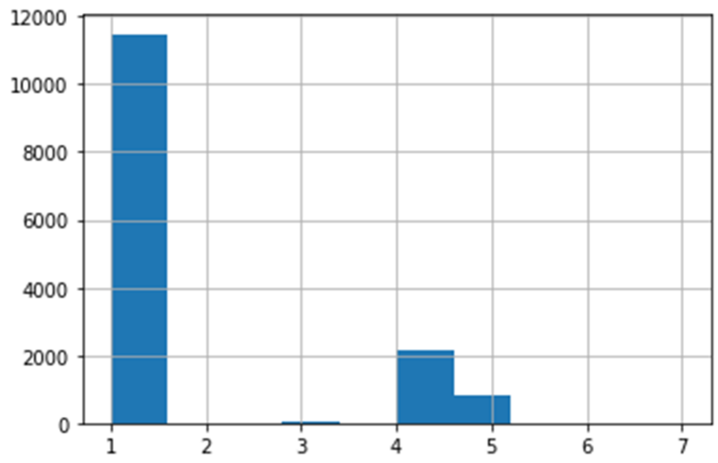

Figure 2.1: Historgram of the test smaples in ”shuttle dataset”

Therefore, Majority of the samples classes are almost at 1, 4 and 5 and minority of samples classes are at (2, 3, 6 and 7 as shown in figure 1.1.
By checking on the percentage of the majority classes compared to the minority, it’s found that it’s 99.6 percentage for majority classes and 0.004 percentage for the minority classes.

Chapter 3

Basic Algorithms

3.1 Logistic Regression[3]

After tuning ($random_state$) parameter ,it's found that accuracy as follows:

$random_state$	Accuracy score	Training $Time(seconds)$	$TestingTime(seconds)$
0	0.9310344827586207	3.6601	0.0462
5	0.9310344827586207	3.7252	0.0053
10	0.9310344827586207	3.9414	0.0061
15	0.9310344827586207	3.7868	0.0050
20	0.9310344827586207	3.9809	0.0047
25	0.9310344827586207	3.6746	0.0046
35	0.9310344827586207	3.5674	0.0045
Mean	0.9310344827586207	3.7430875	0.010125
Standard deviation	0	0.150582018	0.014586075

Table 3.1: Logistic Regression results

3.2 KNN[4]

After tuning $n_neighbors$, it's found that accuracy as follows:

$n_neighbors$	Accuracy score	Training $Time(seconds)$	$TestingTime(seconds)$
1	0.9988275862068966	0.4155	1.2709
2	0.9981379310344828	0.4095	1.5060
3	0.9982758620689656	0.3981	1.6028
4	0.9981379310344828	0.4007	2.0975
5	0.998	0.4125	2.1601
6	0.9977931034482759	0.4100	1.8651
7	0.9978620689655172	0.4144	1.9694
8	0.9977241379310345	0.4110	2.1193
Mean	0.998094828	0.408963	1.8238875
Standard deviation	0.000351536	0.006286025	0.328148675

Table 3.2: KNN results

3.3 SVM

3.3.1 Linear[5]

$random_state$	Accuracy score	Training $Time(seconds)$	$TestingTime(seconds)$
0	0.9476551724137932	29.6740	0.0050
5	0.9226206896551724	29.9606	0.0068
10	0.9129655172413793	29.4087	0.0056
15	0.9430344827586207	30.3286	0.0058
20	0.9191034482758621	29.1228	0.0049
25	0.9215172413793103	29.4188	0.0050
30	0.9140689655172414	30.4512	0.0047
35	0.9488275862068966	29.2560	0.0056
Mean	0.928724138	29.7025875	0.005425
Standard deviation	0.015175996	0.496012553	0.000681909

Table 3.3: SVM Linear results

3.3.2 Non Linear[6]

$random_state$	Accuracy score	Training $Time(seconds)$	$TestingTime(seconds)$
0.01	0.9975862068965518	21.4657	1.3109
0.001	0.9981379310344828	3.8885	0.2364
0.0001	0.9984137931034482	2.1215	1.6028
0.00001	0.9983448275862069	4.0318	0.7787
Mean	0.99812069	7.876875	0.9822
Standard deviation	0.000375107	9.100773651	0.602997496

Table 3.4: SVM $non - Linear$ results

Chapter 4

Comparison

4.1 Discussion

Here is a table to summarize all the results of all machine learning algorithms:

ML Algorithms	Terms	Mean	Standard Deviation
Logistic Regression	Accuracy score	0.9310344827586207	0
	Training time	3.7430875	0.150582018
	Testing time	0.010125	0.014586075
KNN	Accuracy score	0.998094828	0.000351536
	Training time	0.4089625	0.006286025
	Testing time	1.8238875	0.328148675
SVM(Linear)	Accuracy score	0.928724138	0.015175996
	Training time	29.7025875	0.496012553
	Testing time	0.005425	0.000681909
SVM(Non-Linear)	Accuracy score	0.99812069	0.000375107
	Training time	7.876875	9.100773651
	Testing time	0.9822	0.602997496

Table 4.1: Results of all Machine Learning algorithms

4.2 Mean

4.2.1 Accuracy

According to the results shown in table 3.1,figure 3.1 represents the mean of the three machine Learning algorithms in terms of Accuracy:

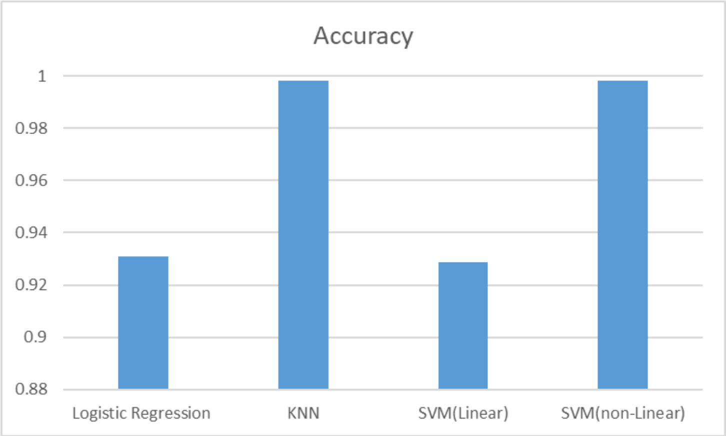

Figure 4.1: Mean of Accuracy of basic algorithms

4.2.2 Training Time

According to the results shown in table 3.1,figure 3.2 represents the mean of the three machine Learning algorithms in terms of training time:

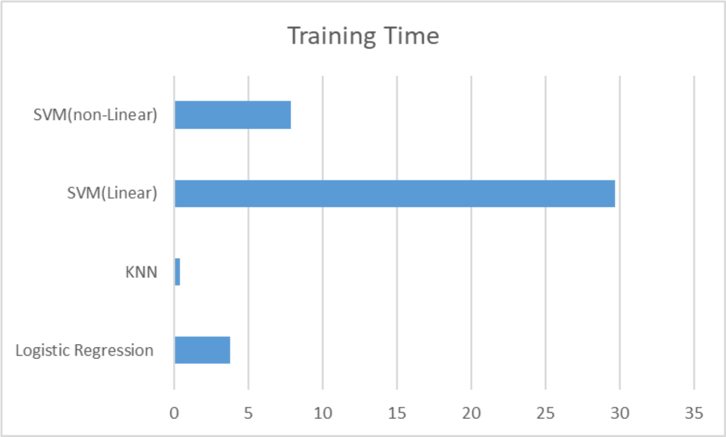

Figure 4.2: Mean of Training Time of basic algorithms

4.2.3 Testing Time

According to the results shown in table 3.1 ,figure 3.3 represents the mean of the three machine Learning algorithms in terms of testing time:

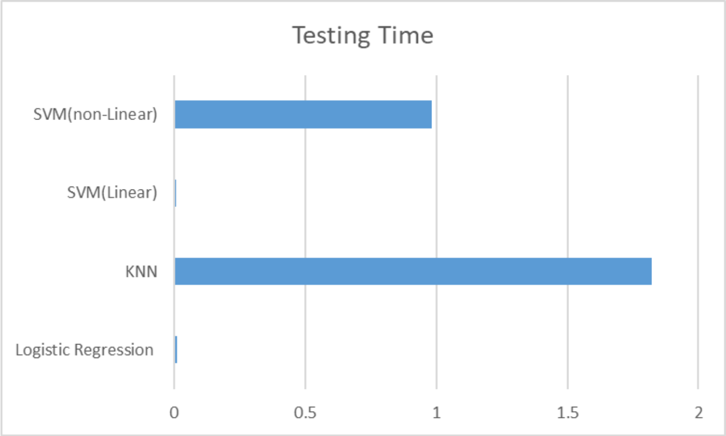

Figure 4.3: Mean of Testing time of basic algorithms

4.3 Standard Deviation

4.3.1 Accuracy

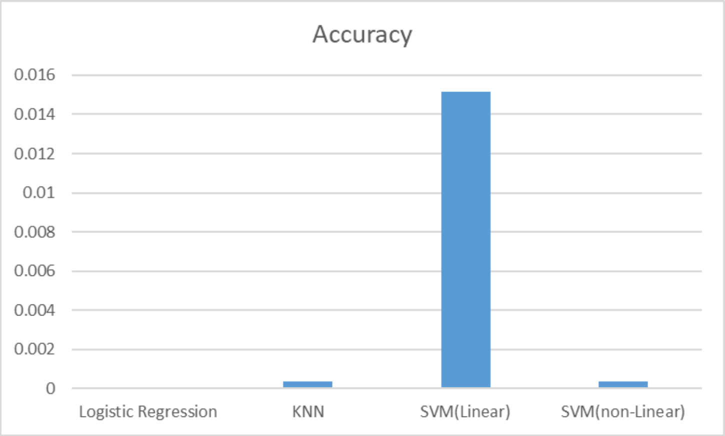

Figure 4.4: Standard deviation of accuracy of basic algorithms

4.3.2 Training Time

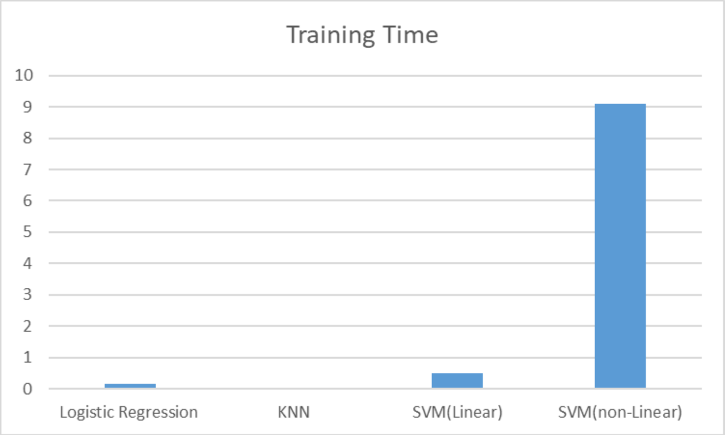

Figure 4.5: Standard deviation of training time of basic algorithms

4.3.3 Testing Time

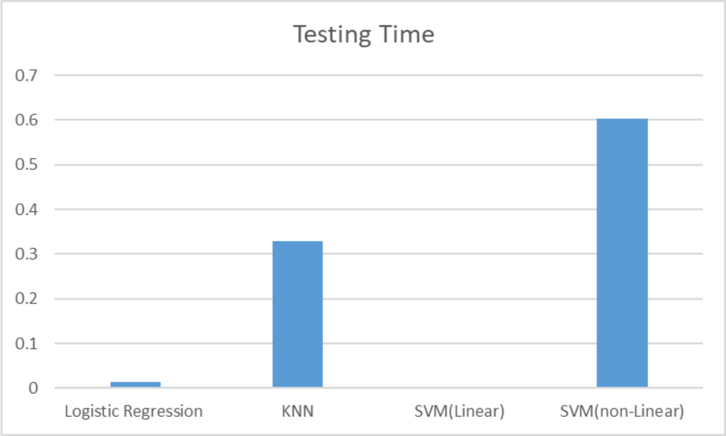

Figure 4.6: Standard deviation of testing time of basic algorithms

Chapter 5

Proposed Machine Learning Algorithm

5.1 Discussion

After trying the other machine learning techinques, it's found that decision tree machine learning techique results are satisfying in terms of accuracy, training time and testing time. Here are the results from the benchmark application:

At $random_statе$=0:
Training time: 0.2061 seconds
Testing time: 1.6885 seconds
Accuracy of the Decision Tree SVM is 0.9975862068965518

At $random_state$=5:
Training time: 0.0834 seconds
Testing time: 1.2920 seconds
Accuracy of the Decision Tree SVM is 0.9975862068965518

According to the above results,decision tree machine learning techique has better results compared to the other algorithms in terms of accuracy, training time and testing time.

Chapter 6

The left and right distribution effects

The regularization parameter lambda, we can then control how well we fit the training data while keeping the weights small. By increasing the value of lambda, we increase the regularization strength.
The parameter C that is implemented for the LogisticRegression class in scikit-learn comes from a convention in support vector machines, and C is directly related to the regularization parameter lambda which is its inverse: $c = 1/lambda$

The requirement of having different values of lambda works for Logistic Regression and SVM.

6.1 Discussion

Logistic Regression

Therefore, after tuning the parameter c the result is as follows:

At c=10**5:
Training time: 5.1428 seconds
Testing time: 0.0057 seconds
Accuracy of the Logistic Regression is 0.9351034482758621

At c=10**-5:
Training time: 1.1438 seconds
Testing time: 0.0044 seconds
Accuracy of the Logistic Regression is 0.9210344827586207

According to the above results, it shows, the higher the value of lambda ,the lower the accuracy of Logistic Regression machine learning algorithm.

SVM

Therefore, after tuning the parameter c the result is as follows:

At c=10**5:
Training time: 26.9597 seconds
Testing time: 0.0043 seconds
Accuracy of the linear SVM is 0.8713793103448276

At c=10**-5:
Training time: 19.6836 seconds
Testing time: 0.0037 seconds
Accuracy of the linear SVM is 0.9215172413793103

According to the above results, it shows, the higher the value of lambda ,the higher the accuracy of SVM machine learning algorithm..

Bibliography

[1] Newman, D.J., Asuncion, A., 2007. *"UCI Machine Learning Repository"* . University of California, Irvine, Dept. of Information and Computer Sciences .

[2] sklearn: Machine Learning in Python,
`https://scikit-learn.org/s`

[3] Dreiseitl, Stephan, and Lucila Ohno-Machado. *"Logistic regression and artificial neural network classification models: a methodology review"*. Journal of biomedical informatics 35, no. 5-6 (2002): 352-359.

[4] Cai, Yun-lei, Duo Ji, and Dongfeng Cai. *"A KNN Research Paper Classification Method Based on Shared Nearest Neighbor."*. In NTCIR, pp. 336-340. 2010.

[5] Tang, Yichuan. *"Deep learning using linear support vector machines."*. arXiv preprint arXiv:1306.0239 (2013)

[6] Chen, Yunqiang, Xiang Sean Zhou, and Thomas S. Huang. *"One-class SVM for learning in image retrieval."*. In Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205), vol. 1, pp. 34-37. IEEE, 2001.

[7] Dietterich, Thomas G., and Eun Bae Kong. *"Machine learning bias, statistical bias, and statistical variance of decision tree algorithms."*. Technical report, Department of Computer Science, Oregon State University, 1995.