# RESULTS AND DISCUSSION

## Metrics

The models have various predictive powers which needs proper measures to evaluate the classifier. We have used accuracy score and F1-score for this.

*4.2.1 Accuracy Score.* A common metric which is the fraction of the samples correctly predicted. For a predicted value of i-th sample i.e.   and   being the respective true value, the fraction of right predictions over   may be defined as :

[4]

The mean and standard deviation of the accuracy of the three algorithms has been stated in the table below.

Table 1: **Mean and deviations of accuracies**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Random Forest | Logistic Regression | SVC |
| Mean (%) | 71.99 | 66.79 | 64.86 |
| Standard Deviation | 1.55 | 1.39 | 1.46 |

*4.2.2 F1-Score.* We selected this metric to strike a balance between precision and recall. For =1, F1 is derived from:

The mean and standard deviation of F1 scores for the three algorithms have been mentioned in the table.

Table 2: **Mean and deviations of F1 score**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Random Forest | Logistic Regression | SVC |
| Mean | 0,70 | 0.63 | 0.58 |
| Standard Deviation | 0.02 | 0.02 | 0.02 |

## Discussion

We ran 20 iterations on the dataset to check how the three algorithms responded with each iteration. For the 0th iteration, the dataset was unpruned and random forest responded the best. We used the metric measures to determine algorithm statistics and performance. For 2nd iteration, the accuracy decreased for random forest while it increased for logistic regression and decreased for SVM; however the rankings remained unchanged. On the other hand, for its 6th iteration, the accuracy for random forest increased by 3% when compared to its original accuracy, logistic regression show negligible increase whereas SVM showed 3% decrease in its accuracy. On the whole, fluctuations were seen in their respective accuracies with each iteration.

## Results

Some related literary works on movie datasets were mostly centred on random forest decision trees while some focused improving SVM accuracy. We ran unbiased analysis on the three algorithms and observed that random forest performed the best followed by logistic regression and SVM. Their rankings remain unchanged on unpruned and pruned datasets across two metric measures. However, several iterations showed some fluctuations in terms of their accuracies. To conclude, pruning of datasets didn’t affect algorithm rankings.