Analysis of Different Parametric and Non-Parametric Statistical tests for Text and Non-Text Classification

A thesis submitted in the partial fulfillment of the requirement for the

Degree of Master of Computer Application

of

Jadavpur University

By

DEBARATI ROY

Registration Number: 137312 of 2016-2017

Examination Roll Number: MCA196004

Under the guidance of

Dr. Nibaran Das

Associate Professor

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

May 2019

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JADAVPUR UNIVERSITY

TO WHOM IT MAY CONCERN

This is to certify that the thesis entitled "APPLICATION OF DEEP LEARNING TECHNIQUE FOR MULTI-SPECTRAL IMAGE CLASSIFICATION" has been satisfactorily completed by Ishita Das (University Registration No.: 137312 of 2016-17, Examination Roll No: MCA196004). It is a bonafide piece of work carried out under my guidance and supervision and be accepted in partial fulfillment of the requirement for the Degree of Master of Computer Science and Engineering, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Jadavpur University, Kolkata.

Dr. Nibaran Das (Thesis Supervisor)
Associate Professor
Department of Computer Science and Engineering
Jadavpur University, Kolkata-700032

Countersigned:

Prof. Mahantapas Kundu
Head, Department of computer Science and Engineering
Jadavpur University, Kol-700032

Prof. Chiranjib Bhattacharjee
Dean, Faculty of Engineering and Technology
Jadavpur University, Kol-700032

FACULTY OF ENGINEERING AND TECHNOLOGY
JADAVPUR UNIVERSITY

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains a literature survey and original research work done by the undersigned candidate, as part of his MCA studies.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material results that are not original to this work.

Name	: DEBARATI ROY
University Registration No.	: 137312 of 2016-17
Examination Roll No.	: MCA196004
Thesis Title	: Analysis of Different Parametric and Non-Parametric Statistical tests for Text and Non-Text Classification

____

Signature with date

JADAVPUR UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE OF APPROVAL

The foregoing thesis is hereby accepted as the credible study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the thesis only for the purpose for which it is submitted.

____

Signature of Examiner

Date:

____

Signature of Supervisor

Date:

ACKNOWLEGEMENT

First and foremost, I would like to start by thanking God Almighty for showering me with the strength, knowledge and potential to embark on this wonderful journey and to persevere and complete the embodied research work satisfactorily.

I am pleased to express my deepest gratitude to my thesis guide, **Dr. Nibaran Das**, Department of Computer Science and Engineering, Jadavpur University, Kolkata for his invaluable guidance, constant encouragement and inspiration during the period of my dissertation.

I am highly indebted to **Jadavpur University** for providing me the opportunity and the required infrastructure to carry on my thesis. I am also grateful to the **Center for Microprocessor Applications for Training Education and Research** for giving me the proper laboratory facilities as and when required. I am thankful for all the teaching and non-teaching staff whose helping hands have smoothed my journey through the period of my research.

Last but not the least; I would like to thank my family members, classmates, seniors and friends for giving me constant encouragement and mental support throughout my work.

Debarati Roy

University Registration No.: 137312 of 2016-17
Examination Roll No.: MCA196004
Master of Computer Application
Department of Computer Science and Engineering
Jadavpur University

CONTENTS

1

CHAPTER ONE

INTRODUCTION

Comparing machine learning methods and selecting a final model is a common operation in applied machine learning. In general, we evaluate models using resampling methods like k-fold cross validation from which we analyze the mean accuracy score. Although this approach can be misleading if the mean skill scores is a result of statistical fluke instead of real. Also “accuracy” is not a good measurement for classification problems as data-distribution can be skewed. For this uncertainty, every analyzer and researcher wants to do significance analysis of their experimented outcomes for that reason Statistical significance tests are designed to address this problem and quantify the likelihood of the samples of various evaluation metrics of classification models. Practical application of statistical significance analysis is Hypothesis-testing. In the language of Hypothetical-testing, if the null hypothesis is rejected, it suggests that there is a significant difference of performances of our selected models applied on a single dataset. In this article, we propose a set of simple, yet safe and robust parametric and non-parametric tests for statistical comparisons of the classifier. As an example of the univariate parametric test, we propose “T-test”, ”Z-test”, “Two-Sample-T-test”,” Two-Sample-Z-test”,” ANOVA-test”,” Fisher-Exact-Test” etc. For the multivariate parametric test, we propose “MANOVA-Test”. For univariate Non-Parametric test, we propose “Chi-Square-test”, ”Median-test”, “McNemar-test”, ”Mann-Whitney-test”,” Two-Sample-Kolmogorov-Smirnov-test”.[1][2]

In the field of machine learning, nowadays the big problem is to evaluate the best classification model for a specific kind of data. We know that there are many kinds of data (As an example –text-type, image-type, etc.) Also, the size of the data manipulates the performance of classification algorithms. Our main goal is to find out the optimal classification model for a particular kind of data. The easiest way to measure the performance of classification algorithms is to calculate the accuracy-scores of the classification techniques on the particular kind of data. But we know that accuracy is not a good measure as data distribution can be skewed. Also, it does not tell the underlying distribution of response values, and it does not tell us what types [1]of errors our classifier is making. As an example, there is a problem which is called “accuracy-paradox” when true-positive-rate is less than false-positive-rate but the accuracy-score is still good.[3]

Let us consider an example,

Table 1Accuracy paradox

	Classified-positive	Classified-negative
Positive-class	0(TP)	25(FN)
Negative-class	0(FP)	125(TN)

Here we get accuracy= (0+125)/ (0+125+0+25) =83.3%

But the model is completely useless one with exactly zero predictive power and yet, we got an increase in accuracy. Though there are many more metrics to evaluate the performance of the different classification algorithms on same data as an example-roc_auc_score, false-positive-rate (fpr), true-positive-rate (tpr), sensitivity, specificity, f_score, etc. But our aim is “significance Analysis” of classification algorithms, for that we have to take help of statistics.

Now the question is that what is the statistical significance?

- Statistical significance is the strength of observation the statistical differences among the variables (two or more).
- Statistical significance refers to the unlikelihood that means differences occurred in the samples due to sampling error.
- Statistical significance is the likelihood that a relationship between two or more variables caused by something other than chance (probability).

Statistical significance level reflects risk-tolerance and confidence level. The most practical use of statistical significance is **Hypothetical-Testing**. Hypothesis test provides p-value that is a probability value. The accepted p-value that makes a hypothesis statistically significant is 0.05(i.e. 5%) in general. Now if our calculated p-value is greater than or equal to significant value, than our hypothesis is true else not.

For Hypothetical-statistical analysis, we have to consider some parameters. We mainly consider the six evaluation metrics such as **True-Positive, True-Negative, Sensitivity, Specificity, Positive-Predictive-Value, and Negative-Predictive-Value.**

CHAPTER TWO

LITERATURE REVIEW

One of the most cited paper in this field is by “Olcay Taner Yıldız1, ¨Ozlem Aslan2, and Ethem Alpaydın” named “Multivariate Statistical Tests for Comparing Classification Algorithms [2008]”. Their main aim on this paper is to discriminate between the univariate and multivariate hypothetical tests on the case of measurement the performances of classification algorithms on a single dataset. They have used mainly four tests, for univariate statistical tests they have used T-test and ANOVA-test and for multivariate statistical tests they have used Hotelling’s-T-Squared-test and MANOVA test. They have mainly focused on bioinformatics datasets. They claim that they have used 36 kinds of different datasets. For univariate statistical analysis, they have used the sum of true-positive-value and false-positive value of different classifiers as a parameter. For multivariate statistical tests, they have considered the two pairs of evaluation metrics such as (true-positive, false-positive) and (sensitivity, specificity).[4]

We have analyzed the paper on different kind of datasets.

ANOVA vs. MANOVA

On digit dataset, we have applied a number of classification algorithms for this analysis we consider mainly three classifier model i.e. Logistic-Regression, Decision-Tree, and KNN. We have used the sum of true-positive-value and false-positive-value as a parameter for the “ANOVA” test.

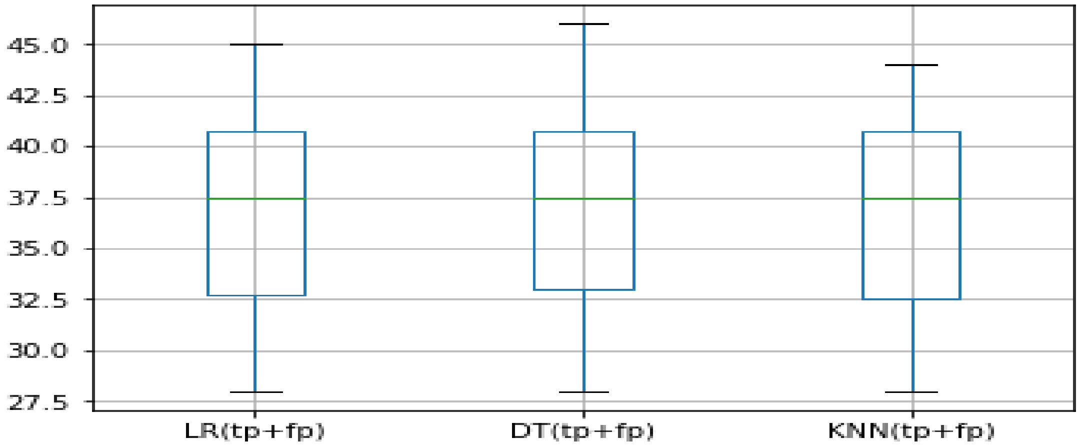


Figure 1 boxplot for digit dataset(tp+fp)

In this boxplot diagram, we can observe that there is no difference between the means of the parameter (tp+fp) of the classifier models. So generally it seems that there is no difference between the performances of the classifiers.

Now the result of ANOVA analysis is,

Table 2 ANOVA table for digit

F-Statistics	**P-value**
0.0	**1.0**

On the basis of the 5% confidence interval, the p-value got from the ANOVA test is 1.0, that is greater than 0.05(5%) that means the **Null Hypothesis** is accepted.

Now for MANOVA test on the same classifier models' performance, we have considered the two parameters true-positive(TP) and false-positive(FP).

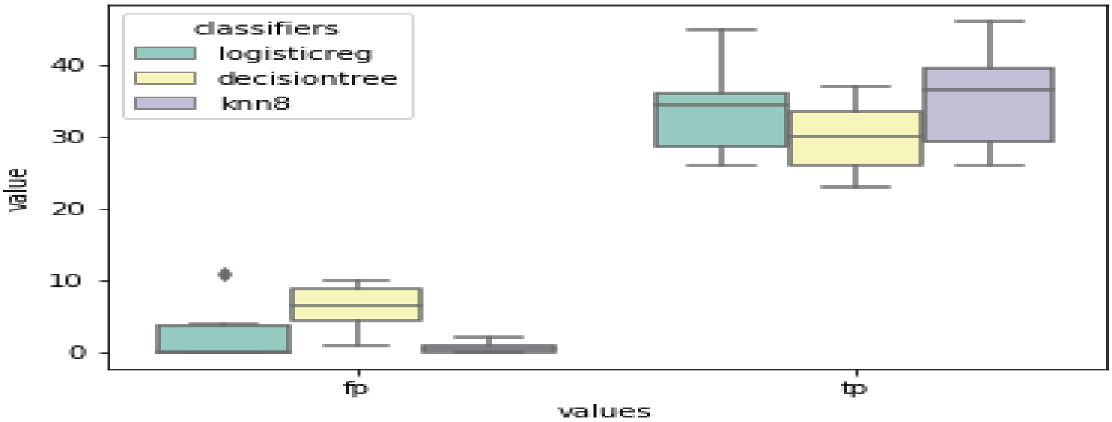


Figure 2 boxplot diagram for tp and fp on digit dataset

We can observe on the boxplot diagram there is a huge difference among false-positive-values as well as true-positive-values of the classifier models "Logistic-Regression", "Decision-Tree" and "KNN" which we could not find out with the help of "ANOVA". For that reason, we need multivariate analysis that can be done by "MANOVA" test.

The repeated measure of MANOVA is,

Table 3 MANOVA table

	sum_sq	**df**	**F**	**p-value**
C(values)	1.338027e+04	1.0	6.318069e+02	1.770362e-31
C(classifiers)	5.863866e-28	2.0	1.384438e-29	1.000000e+00
C(values):C(classifiers)	3.361333e+02	2.0	7.935992e+00	9.515873e-04
Residual	1.143600e+03	54.0	NaN	NaN

We can see that the marked p-values are less than .05.

The P-value obtained from MANOVA analysis for values, classifiers, and interaction is statistically significant (P<0.05). We conclude that the type of values significantly affects the yield outcome, different classifier algorithms significantly affect the yield outcome, and the interaction of both values and classifiers algorithm significantly affects the yield outcome

so we reject the null hypothesis

CHAPTER THREE

HYPOTHESIS TESTING

Hypothesis testing is a practical approach to significance analysis of data. The whole structure of hypothesis test depends on the assumption of two mutually exclusive statements about the examined population to determine which statement is best supported by the examined data. When we say that our evaluated result is statistically significant, it’s thanks to a hypothesis test.[4][5]

Hypothesis-Test is of two types.

- Parametric Test
- Non-Parametric Test

Now the main differences between Parametric and Non-Parametric tests are,

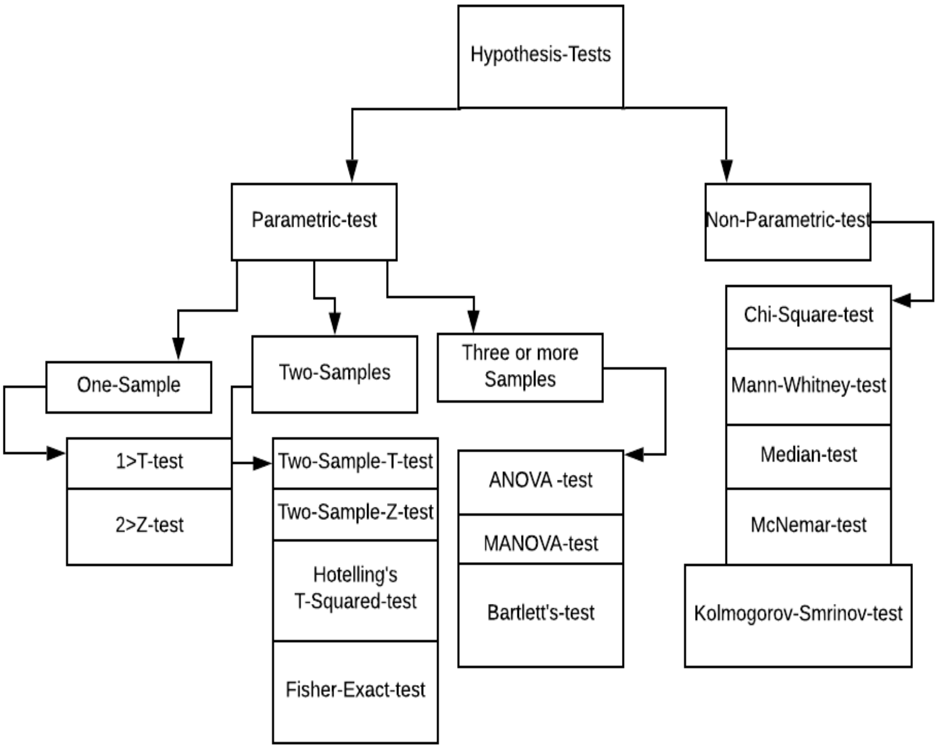
Parametric-Test	Non-Parametric-Test
1>These are those statistical tests where the information about the population is completely known by means and ways of its parameters.	1>These are those statistical tests where there is no knowledge about the population and parameters but still it is required to test the hypothesis of the population.

2> It assumes that the data comes from a type of probability distribution and then it makes inferences about the parameters of the distribution.	2>It covers all the techniques that do not rely on data belonging to any particular distribution.
3> It makes more assumptions. If they are correct, the test can produce accurate p-value otherwise the test could be misleading.	3>It makes very few assumptions, therefore, its applicability is much wider.

In this article, we have tried some Parametric as well as some Non-Parametric statistical tests for significance analysis of the performance of the classical models.

Here we represent a diagram in the graphical format of all the statistical tests which we have covered.

Figure 3 Different Hypothesis tests



3.1 ONE-SAMPLE PARAMETRIC-TEST

For any classification model, our main aim is to increase the true-positive-rate (TPR). So to analyze the performance of two classification models on the same data we follow the path.

Let A and B be two classification models and tpAi and tpBi are the true positive values for model A and B respectively.

Now we consider a variable E such that,

Equation 1

$$E_i = (tp_{Ai} - tp_{Bi}) \quad [i=1, 2, 3\ldots.n \text{ (n is sample size)}]$$

Now on the variable E we apply One-Sample-Parametric test.

In a similar technique, we can analyze the other evaluation metrics such as true-negative, sensitivity, specificity, PPV, NPV.

T-test: - The one sample t-test is a statistical test which is used to determine whether a sample of observations has the desired mean or not at a significance level. In this case, our hypothetical assumption is

The null hypothesis (H0) assumes that the difference between the computed mean (μ) and the comparison value (m0) is equal to zero.

The alternative hypothesis (H1) assumes that the difference between the computed mean (μ) and the comparison value (m0) is not equal to zero.

Here we consider that m0=0. That means if null-hypothesis is accepted for 5% confidence-interval then there is no difference between the true-positive-rate of the classification models A and B. For one sample to test the sample size should be less than 30 for the sake of accuracy.[5]

Z-test: - The test statistics of one-sample-Z-test is as same as one-sample-T-test. But Z-test is used in place of T-test when the sample size is greater than 30.

3.2 TWO-SAMPLE-PARAMETRIC-TEST

To analyze the performance of two classifiers on the same data, we mainly take help of two-sample-hypothesis test parametric as well as non-parametric. At first, we will discuss the two-sample-parametric tests. For the two-sample-parametric test, we mainly consider the evaluation metrics such as,

True-Positive which is an outcome where the classifier model correctly predicts the positive class.

Sensitivity which is a measure of the proportion of actual positive cases that got predicted as positive.

Sensitivity=True-positive/ (True-positive+False-negative)

True-Negative which is an outcome where the classifier model correctly predicts the positive class.

Specificity which is a measure of the proportion of actual negative cases got predicted as negative.

Specificity=True-negative/ (True-negative+False-positive)

Positive-Predictive-value which is the proportion of predicted positives which are actual positives.

PPV=True-positive/ (True-positive+False-positive)

Negative-Predictive-value which is the proportion of predicted negatives which are actual negatives.

NPV=True-negative/ (True-negative+False-negative)

But for analysis for Fisher's-Exact-test, Chi-Square-test and McNemar tests we mainly consider the predicted values coming from the selected two classifiers.

Now let's elaborate the all two-sample-parametric tests which we have applied on the metrics mentioned above.

3.2.1. TWO-SAMPLE-T-TEST

The two-sample t-test mainly compares the means of two independent groups in order to determine whether there is statistical proof that the two independent samples population means are significantly different or not. The Independent Samples t-test is a parametric test.

Equation 2

Formula = |X1-X2|/Sqrt ((S1^2/n1) + (S2^2/n2))

(X1=mean of sample 1. X2=mean of sample 2. S1=variance of sample 1. S2=variance of sample 2. n1=size of sample 1. n2= size of sample 2.)
The t-test is applied when

1) Population Standard deviation is not available.

2) n should be less than 30 for better accuracy.[6]

3.2.2. TWO-SAMPLE-Z-TEST

The test statistics of two sample z test is as same as the two-sample t-test.

We would use a Z test if:

- If the sample size is greater than 30, we should use Z-test instead of t-test.
- Data points should not be dependent on each other. In other words, one data point isn’t related or doesn’t affect another data point.
- Data should be normally distributed. But for large sample sizes (over 30) this doesn’t always matter.
- Sample sizes should be equal if at all possible.

3.2.3. HOTELLING’S-T-SQUARED-TEST

Hotelling’s T-Squared (Hotelling, 1931) is the multivariate version of the T-test. “Multivariate” means that we have data for more than one parameter for each sample.

Hotelling’s T-Squared is based on Hotelling’s T2 distribution and forms the basis for various multivariate.[5]

Hotelling’s T-squared has several advantages over the t-test:

- The Type I error rate is well controlled,
- The relationship between multiple parameters is the main concern.
- The test can generate an overall conclusion even if multiple two-sample-tests are inconsistent. While a t-test will tell us which variable is differences among the

groups, Hotelling’s-T-Squared-test summarizes the differences between every pair of groups.

The test hypotheses are:

- Null hypothesis (H0): the two samples have the same multivariate mean.
- The alternate hypothesis (H1): the two samples have different multivariate means.

Three major assumptions are that the samples:

- The two samples should be in normal distributions.
- The two samples should be independent.
- The two samples should have equal variance-covariance matrices (for the two sample test only).

In the case of Hotelling’s-T-Squared-test, we have used the pair of evaluation metrics such as as-(TP, FP), (Sensitivity, Specificity), (PPV, NPV) of the classification model.

Now the test statistic is

Equation 3

$$F=(n_1+n_2-p-1)/p(n_1+n_2-2)\ T^2 \sim F_{p,}\ n_1+n_2-p-1$$

Where:

- n1 = size of sample1,
- n2 = size of sample2,
- p = number of parameters measured,
- n1 + n2 – p – 1 = degrees of freedom.

3.2.4. FISHER’S-EXACT-TEST[6]

Fisher’s-Exact-test is based on the hypergeometric distribution. The test acts on 2X2 contingency table.

There are certain terminologies that help in understanding the theory of Fisher’s exact test. Fisher’s-Exact-test computes p-value according to the hypergeometric distribution of data using binomial coefficients.

The Fisher Exact test uses the following formula:

Equation 4

$$p= ((a + b)!\ (c + d)!\ (a + c)!\ (b + d)!\ ) \ /\ a!\ b!\ c!\ d!\ N!$$

In this formula, the ‘a,’ ‘b,’ ‘c’, ‘d’ are the individual frequencies of the 2X2 contingency table, and ‘N’ is the total frequency.

But there is a conflict that Fisher’s-Exact-test is parametric or non-parametric. We assume that Fisher's exact test is parametric as the test assumes an underlying binomial distribution for the 2X2 contingency table.

As the test acts on 2X2 contingency table only so the test is applicable only for binary classification problems.

3.3. THREE-OR-MORE-SAMPLES-PARAMETRIC-TESTS

When we apply three or more than three classifier models on a single dataset than for significance analysis of the performances of the classifiers we use this kind of parametric tests. Like two-sample-parametric tests, we have used the same evaluation metrics for this kind of parametric tests.

Now let’s elaborate the all three-or-more-sample-parametric tests which we have applied for performance analysis of the classifiers.

3.3.1. ANOVA TEST

ANOVA is a statistical method used to compare the means among three or more groups.

Equation 5

F = between group variability / within group variability.

ANOVA test follows the F-distribution.

Unlike the z and t-distributions, the F-distribution does not have any negative values because within-group, variability is always positive as each distribution getting squared.

The one-way ANOVA compares the means between every pair of two groups and determines whether any of those means are significantly different from each other. Specifically, it tests the null hypothesis:

H0: μ0=μ1=μ2=………=μk[7]

Where μ = group mean and k = number of groups. If, however, the one-way ANOVA returns a statistically significant result, we accept the alternative hypothesis (HA), which is that there is at least one group whose mean is statistically significantly different from other groups.

Calculate Test statistic: variance ratio, F.

Like two-sample-T-test ANOVA also assumes some assumptions on the samples

- The populations should be in normal distribution from which the samples are drawn.
- The samples should be independent.

3.3.2. MANOVA TEST

The MANOVA (multivariate analysis of variance) is a type of multivariate analysis used to analyze data that involves more than one dependent variable at a time. MANOVA allows us to test hypotheses regarding the effect of one or more independent variables on two or more dependent variables.

MANOVA can detect patterns between multiple dependent variables. But, what does that mean exactly? Let’s work through an example that compares ANOVA to MANOVA.

Suppose we are studying three different classification methods of the supervised machine learning algorithm. This variable is our independent variable. We also have false-positive values and true-positive values. These variables are our dependent variables. We want to determine whether the mean scores for false-positive and true-positive differ between the three classification methods.

MANOVA test statistics =>

Pillai – This is Pillai's Trace, one of the four multivariate criteria test statistics used in MANOVA. We can calculate Pillai's trace using generated eigenvalues. Divide each eigenvalue by (1+eigenvalue).

Hotelling's-This is Lawley-Hotelling's Trace. It is very similar to Pillai's Trace. It is the sum of the eigenvalues and is a generalization of the F statistic in ANOVA.

Wilks – This is Wilk's Lamda. This can be interpreted as the proportion of the variance in the outcomes that are not explained by an effect. To calculate Wilk's Lambda for each eigenvalue, calculate 1/ (1+eigenvalue), then find the product of these ratios.

Roys – This is Roy's largest Root. We can calculate this value by dividing the largest eigenvalue by (1+largest eigenvalue).

3.3.3. BARTLETT'S TEST

There are two types of Bartlett's test.

- Bartlett's test for homogeneity of variances.
- Bartlett's test for sphericity(correlation matrix has an identity matrix).

For significance analysis, we mainly focus on Bartlett's test for homogeneity of variances.

Bartlett's test for homogeneity of variances is used to test that variances are equal for all samples or not. We can apply Bartlett's test on more than 3 samples. It's used when we are sure about that all the samples come from a normal distribution.[8]

the null hypothesis of the test is that the variances are equal for all samples. In statistics terms,

H:$\sigma_1^2=\sigma_2^2=\ldots=\sigma_k^2$.

The alternate hypothesis is that the variances are not equal for one pair or more:

Ha: $\sigma_1^2 \neq \sigma_2^2 \neq\ldots \neq \sigma_k^2$. (k=number of samples).

3.4. NON-PARAMETRIC-TESTS

Nonparametric tests are more robust than parametric tests.

Here we are giving some reasons for the robustness of nonparametric tests over parametric tests.

- Nonparametric tests are used to analyze ordinal or nominal data with small sample size.
- Unlike parametric tests, nonparametric tests do not require making any assumptions about the distribution of the population.
- Nonparametric tests are used when the data is quantitative and has unknown distribution, or the sample size is so small that central limit theorem cannot be applied.[9]

Like parametric tests, in case of nonparametric tests, we have used the same evaluation metrics of the classifiers for significance analysis of performances of the classifiers.

Now let's elaborate the all nonparametric tests which we have applied for performance analysis of the classifiers.

3.4.1. CHI-SQUARE-TEST

Chi-square test is a nonparametric test. The chi-square-test uses frequency data to generate statistics. The test is mainly used for testing independence and goodness of fit. Testing independence determines whether the two observations coming from different populations are dependent. Testing of the goodness of fit determines if the observed frequency distribution matches a theoretical frequency. Chi-square test mainly operates upon contingency table. The contingency table can be any size of a square matrix. For that reason, we can apply the chi-square test on binary classification problem as well as a multiclass classification problem.[7]

This is the two-sample test, to create the contingency matrix we mainly give predicted values as input generated by two different classifier model applied on the same dataset.

Equation 6

$$\text{Chi-square}=\sum(O-E)^2/E$$

O = the Observed value.

E = the Expected value.

Chi-Square test vs. Fisher-Exact test

- Both tests do the same job-comparing proportions.
- The Chi-Square test is an approximate test, so does not give accurate p-value.
- The Fisher-Exact test always gives the correct p-value, but it is computationally very complex.

3.4.2. MANN-WHITNEY-TEST

Mann-Whitney U test is the non-parametric test which is an alternative test to the two-sample t-test. It is used to test whether the means of our selected samples are equal or not. Generally, the Mann-Whitney U test is used when the samples are an ordinal type or when the assumptions of the two-sample-test are not satisfied.[8]

Mann-Whitney U test is a non-parametric test, so it does not assume any assumptions related to the distribution of populations from where the samples are come from.

But some assumptions are also assumed Mann-Whitney U test,

- The populations from which the samples are drawn should be random.
- The two samples should be independent.

The Mann-Whitney U is given by,

Equation 7

$$U_1=R_1-n_1\ (n_1+1)/2$$

Where n1 is the size of sample 1, and R1 is the sum of the ranks in sample 1.

Equation 8

$$U_2=R_2-n_2\ (n_2+1)/2$$

Similarly, n2 is the size of sample2, and R1 is the sum of ranks in sample2.

The smaller value of U1 and U2 is the one used when consulting significance tables. The sum of the two values is given by

Knowing that R1 + R2 = N (N + 1)/2 and N = n1 + n2,

From the above relations, we can find,

U1 + U2 = n1n2.

3.4.3. MEDIAN-TEST

The median test is a non-parametric test. It is a special case of PEARSON'S Chi-Squared-test. that is used to test whether two (or more) independent groups differ in central tendency – that means it tests whether the medians of two or more groups are equal or not. The test calculates a range of values that is likely to include the difference between population medians.

Median test also assumes some assumptions,[3]

- Samples should include only one categorical factor.
- Sample data do not need to be normally distributed.
- The process is better represented by median or sample size is less than 20 observations.
- Sample sizes can be unequal.

Let there are k samples with n_1, n_2,…, n_o observations, calculate the grand median of all $n_1 + n_2 + ... + n_o$ samples. Then 2xk contingency table is created. The chi-square test for independence can then be applied to this table. More specifically

H0: All k populations have the same median

Ha: All least two of the populations have different medians

The median test is very robust against outliers and fairly robust against differences in the shapes of the distributions. The median test has poor performance for normally distributed data, even worse power for short-tailed distributions, but good relative power for heavy-tailed (outlier-rich) distributions.

Median test and Mann-Whitney test follow the same statistics but the main difference is that the Mann-Whitney test acts on only two samples. But the Median test can act on two samples or more than two samples. So Median test is more generic than the Mann-Whitney test.

3.4.4. McNemar Test

The McNemar test is a non-parametric test to compare categorical variables of related samples.

The McNemar test operates upon a contingency table.

A contingency table contains counts of items. Item counts are contingent (mutually exclusive). The McNemar test acts on 2X2 contingency table.

The McNemar test checks the disagreements between two cases match. Therefore, the McNemar's test is a homogeneity test for contingency tables.

In terms of comparing two classification algorithms, the test is commenting on whether the two models agree in the same way or not. It does not measure accurately. This is clear when we look at how the statistic is calculated.

McNemar’s test statistic is calculated as:

statistic = (Yes/No - No/Yes)^2 / (Yes/No + No/Yes)

Where Yes/No is the count of test instances that Classifier1 got positive and Classifier2 got negative, and No/Yes is the count of test instances that Classifier1 got negative and Classifier2 got positive.

- McNemar test follows the Chi-square statistics but the test only acts on a 2X2 contingency table, so By McNemar test, we can only analyze the performances of classifiers on the binary dataset. So Chi-square test is more generic than the McNemar test.
- Again Fisher-Exact test follows the same statistics as the McNemar test. But the difference is, Fisher-Exact test acts well when the sample size is small(<20). The Fisher-exact test assumes an underlying binomial distribution for the 2X2 contingency table, but McNemar does not assume any distribution for the contingency table. Hence McNemar test is more robust than Fisher-Exact-test.

3.4.5. KOLMOGOROV-SMIRNOV-TEST

This is a two-sample test when the samples are independent. Kolmogorov-Smirnov statistics is one of the most commonly used measures to assess the predictive power of a statistical model. This is the measure of the degree of separation between positive and negative distribution.

KS statistics is used to understand how well the model is predicting cumulative good and cumulative bad. The two-sample K–S test is one of the most useful nonparametric methods for comparing two samples, as it is sensitive to differentiate between cumulative distribution functions of the two samples.[12]

Where –

Equation 9

$$D = \text{Maximum}\left|Fn1(X) - Fn2(X)\right|$$

- n1 = Observations from sample 1.
- n2= Observations from sample 2.

F stands for cumulative frequency for each sample.

It has been seen that when the cumulative distributions show large maximum deviation |D|

IT IS INDICATING A DIFFERENCE BETWEEN THE TWO SAMPLE DISTRIBUTIONS.

CHAPTER FOUR

CLASSIFIERS

Throughout the whole experimental process, we have mainly focused on supervised classification techniques.

Supervised Classification:- In mathematical term Supervised learning is where we have input variables (x) and an output variable (Y) and we use an algorithm to learn the mapping function from the input to the output Y = f(X).X stands for feature values and Y stands for target values. The goal is to approximate the mapping function so well that when we have new input data (x) that we can predict the output variables (Y) for that data.

For supervised learning, we always use labeled that for training as well as testing.

Now let us elaborate about all the supervised classifier algorithms, which we have covered.

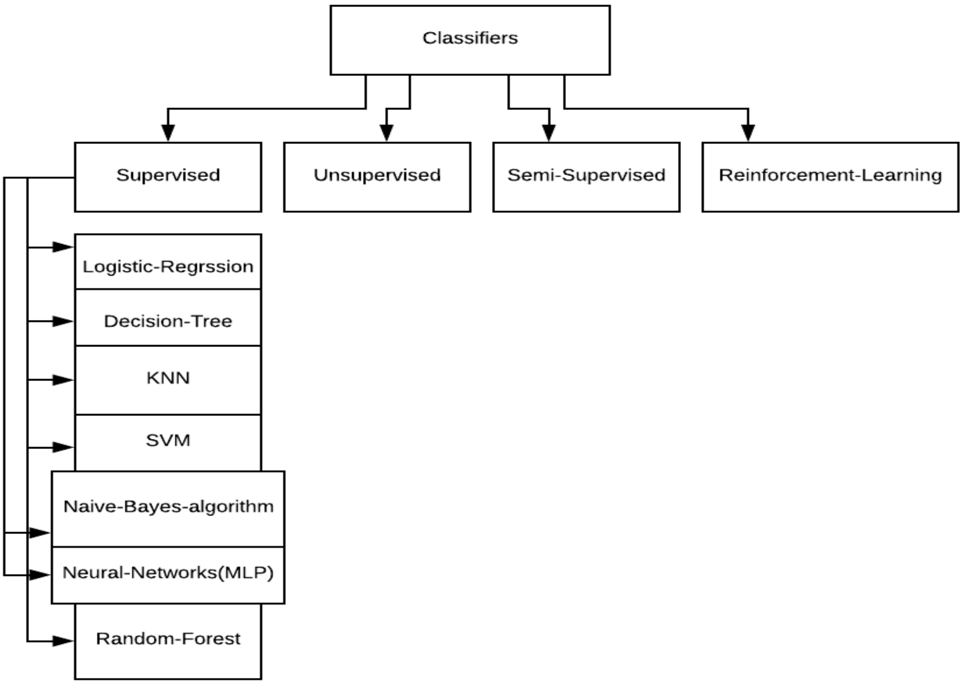


Figure 4 various classifiers

4.1. DECISION-TREE

Decision Tree algorithm makes a decision with the tree-like model. It splits the dataset into two or more homogeneous sets (leaves) based on the most significant differentiators in input variables. To choose a differentiator (predictor), the algorithm considers all features and does a binary split on them (for categorical data, split by category; for continuous, pick a cut-off threshold). It will then choose the one with the least cost (i.e. highest accuracy), and repeats recursively until it successfully splits the data in all leaves (or reaches the maximum depth).

So now our question is that, when does the process terminate?

- Either it has divided into classes that are pure.

Now again a question raised, how can we measure the impurity?

There is a measurement, which is called “entropy”.

Entropy: - Entropy is a degree of randomness of elements or in other words, it is a measure of impurity. Mathematically, it can be calculated with the help of the probability of the items as:

Equation 10

$$H= -\sum p(x) \log p(x)$$

P(x) is the probability of item x. It is a negative summation of probability times the log of the probability of item x.

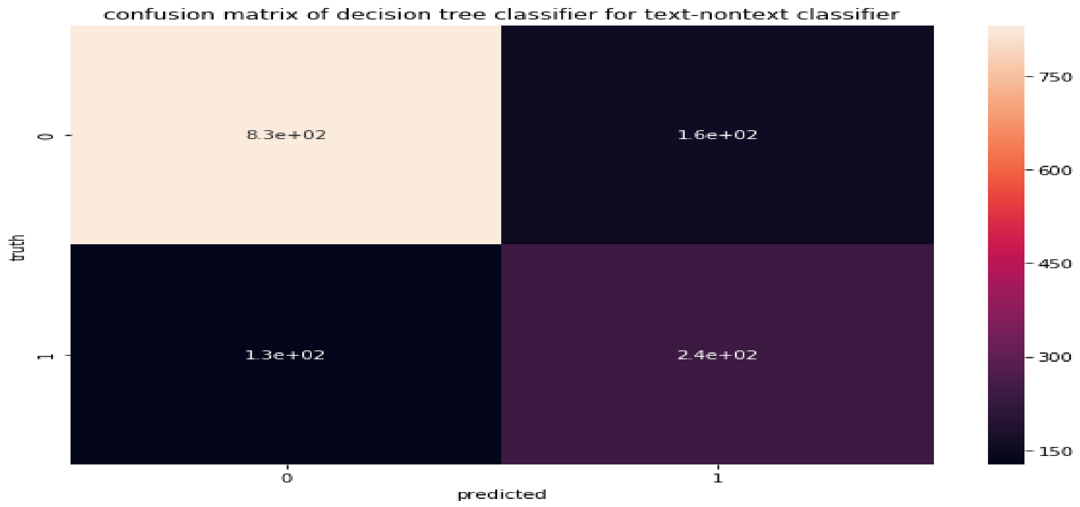


Figure 5 confusion matrix 1

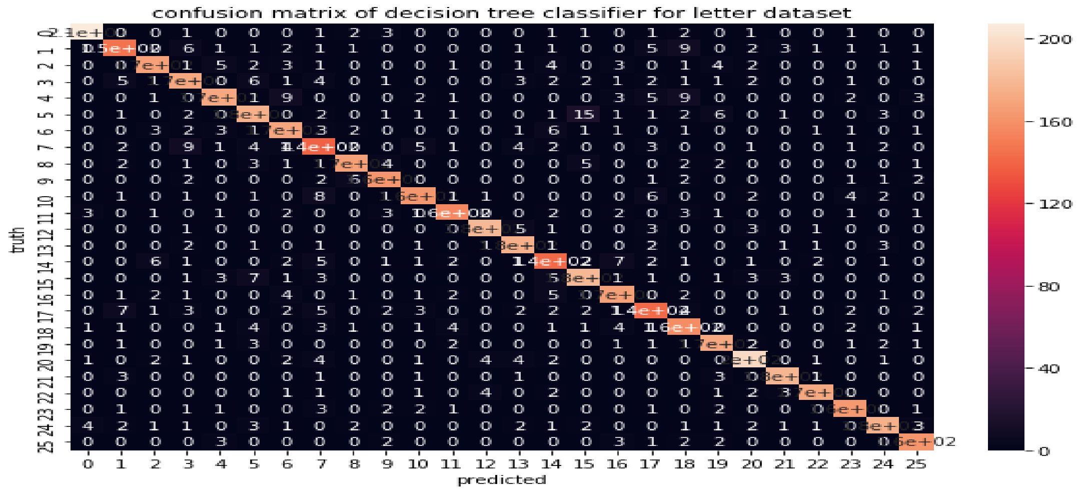


Figure 6 confusion matrix 2

4.2. Random-Forest

It is a tree-based technique that uses a high number of decision trees built out of randomly selected sets of features. Contrary to the simple decision tree, it is highly uninterpretable but it’s generally good performance makes it a popular algorithm.

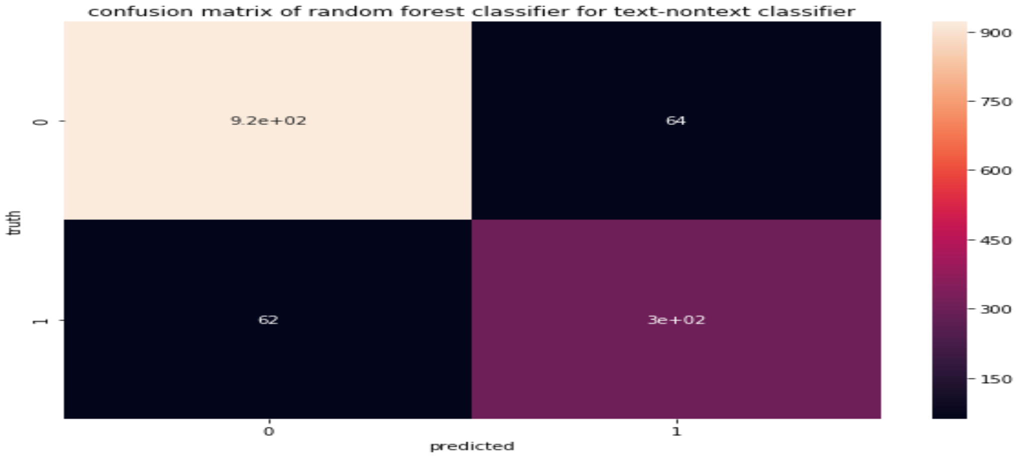


Figure 7 confusion matric 3

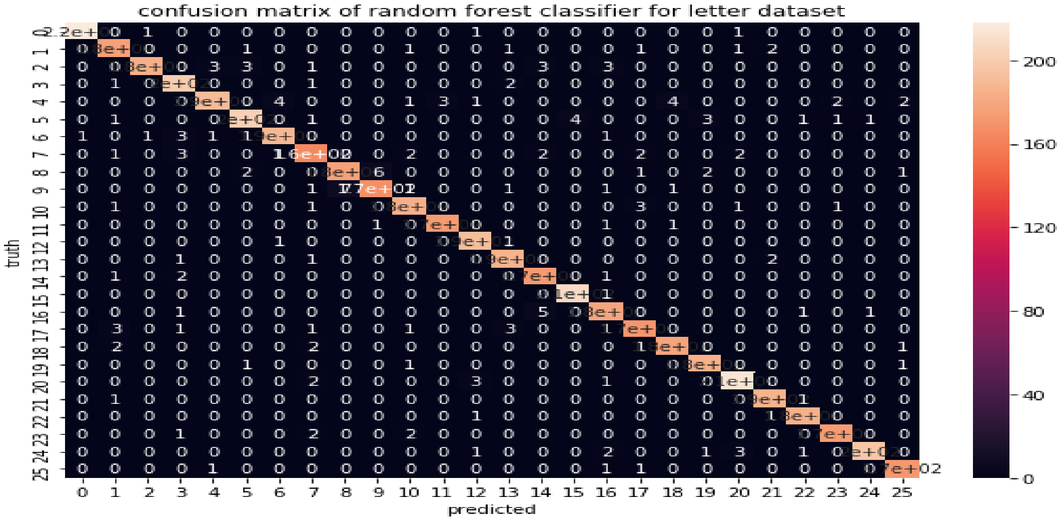


Figure 8 confusion matrix 4

4.3. K-NEAREST-NEIGHBOR

KNN classifies an object by a majority vote of the object's neighbors, in the space of input parameter. The object is assigned to the class which is most common among its k (an integer specifies the number of neighbors is made) nearest neighbor.

It is a non-parametric, lazy algorithm. It's non-parametric since it does not make any assumption on data distribution (the data does not have to be normally distributed). It is lazy since it does not really learn any model and make generalization of the data (It does not train some parameters of some function where input X gives output y). This algorithm mainly follows root-mean-square distance.

So this is not really a learning algorithm. It simply classifies objects based on feature similarity (feature = input variables).

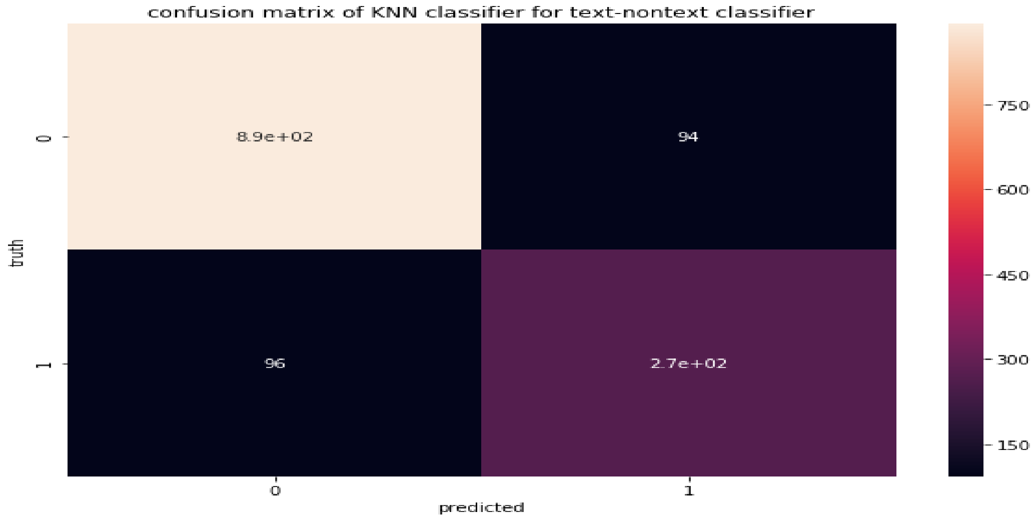


Figure 9 confusion matrix 5

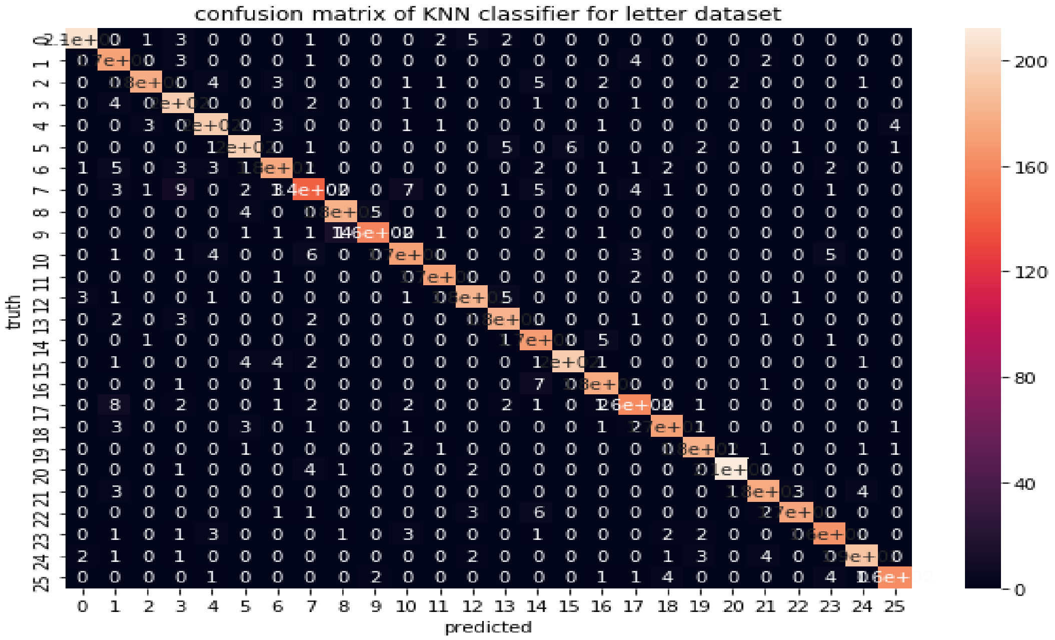


Figure 10 confusion matrix 6

4..4. Logistic-Regression

Logistic regression is a classification algorithm used to assign observations to a discrete set of values. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete values. For that reason, logistic regression is used for classification problems and linear regression is used for regression problems.

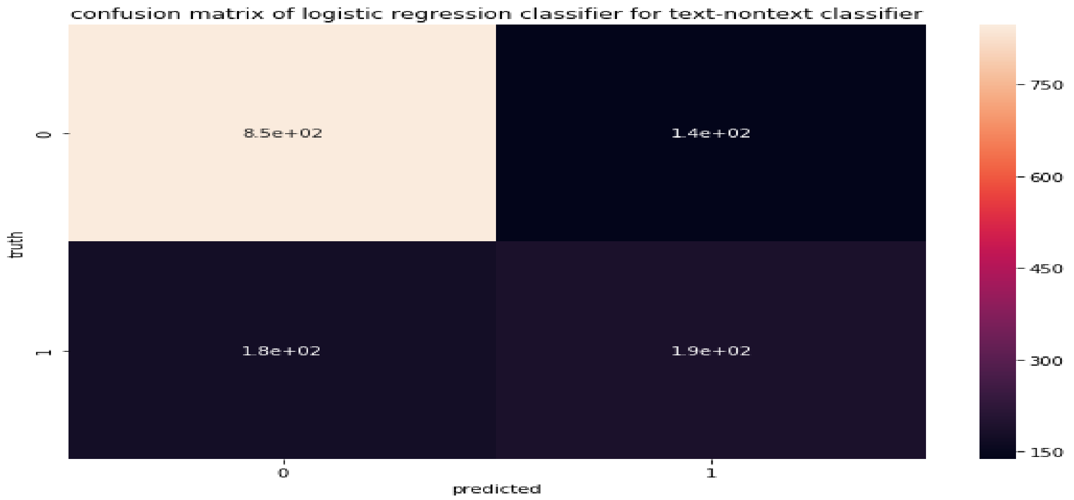


Figure 11 confusion matrix 7

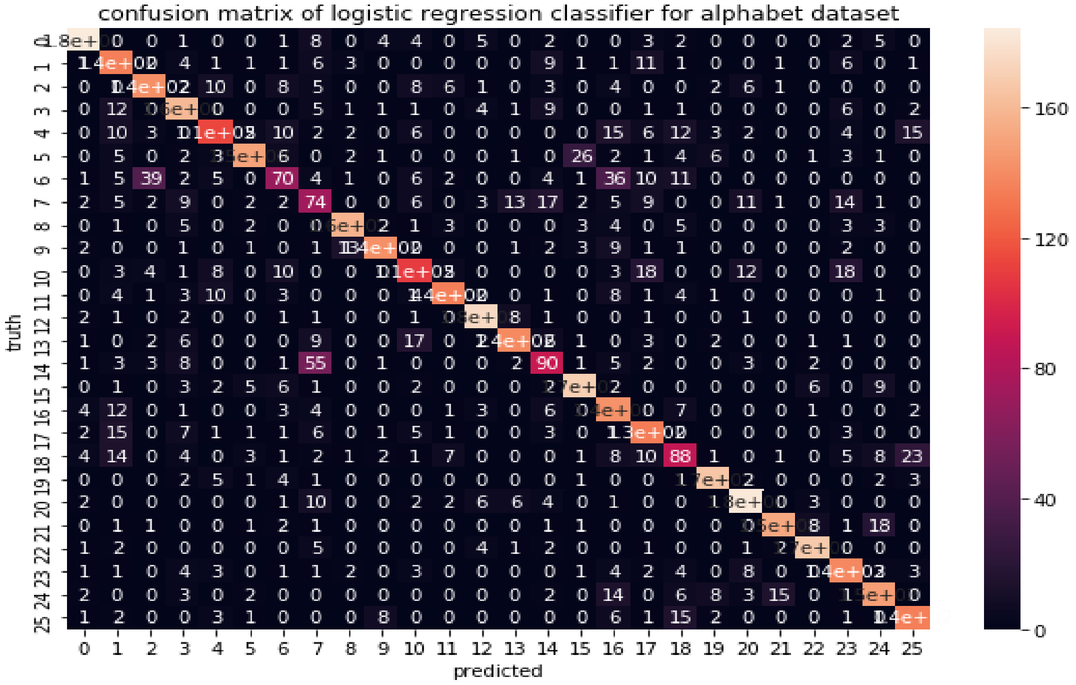


Figure 12 confusion matrix 8

4.5. Support Vector Machine

Support vector machine is a discriminative classifier defined by separating hyperplane. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. So for non-linear classification Support vector machine is ideal.

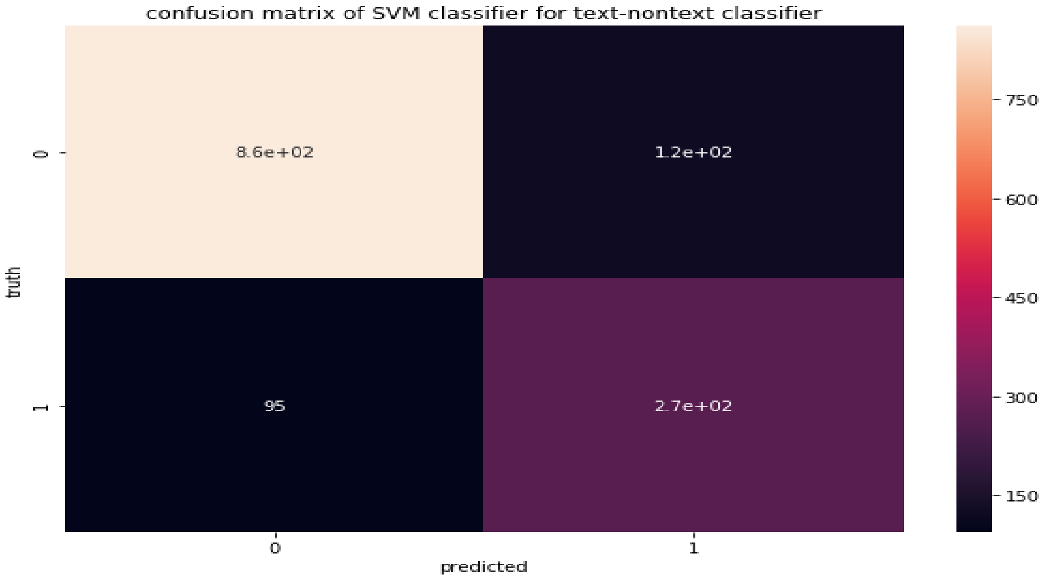


Figure 13 confusion matrix 9

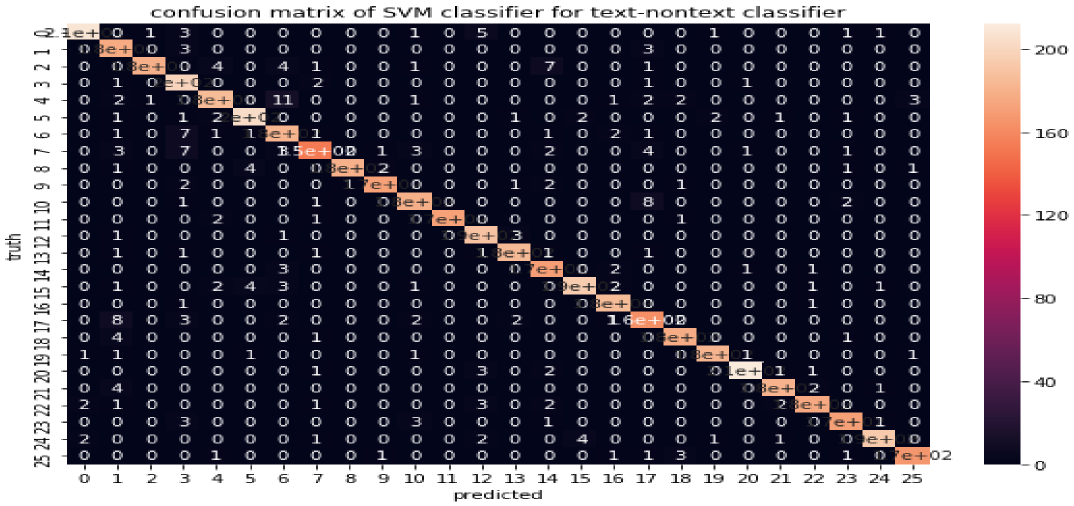


Figure 14 confusion matrix 10

4.6. MULTILAYER-PERCEPTRON

A Multi-Layer Perceptron (MLP) or Multi-Layer Neural Network contains one or more hidden layers (apart from one input and one output layer). While a single layer perceptron can only learn linear functions, a multi-layer perceptron can also learn non – linear functions.

Now let’s describe the actual function of MLP in details.

Let us consider a neural network which takes as input x1, x2, x3 (and a +1 bias term), and outputs f (summed inputs+bias), where f (.) called the activation function. Every activation function (or non-linearity) takes a single number and performs a certain fixed mathematical operation on it. There are several activations.

Throughout the whole experimental we have covered three activation functions for MLP classifier.

- **Sigmoid**: takes real-valued input and squashes it to range between 0 and 1.
- **Tanh**: takes real-valued input and squashes it to the range [-1, 1].
- **ReLu**: ReLu stands for Rectified Linear Units. It takes real-valued input and thresholds it to 0 (replaces negative values to 0).

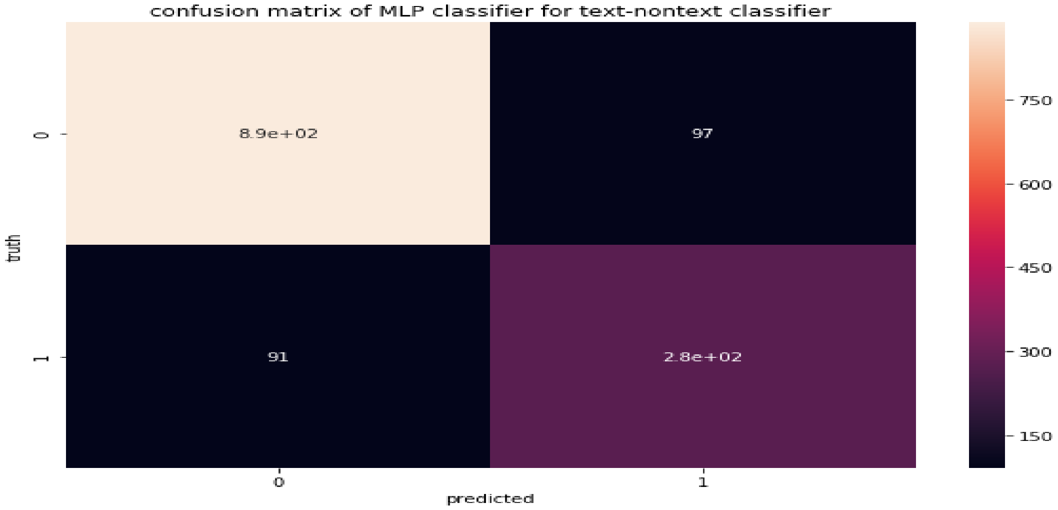


Figure 15 confusion matrix 11

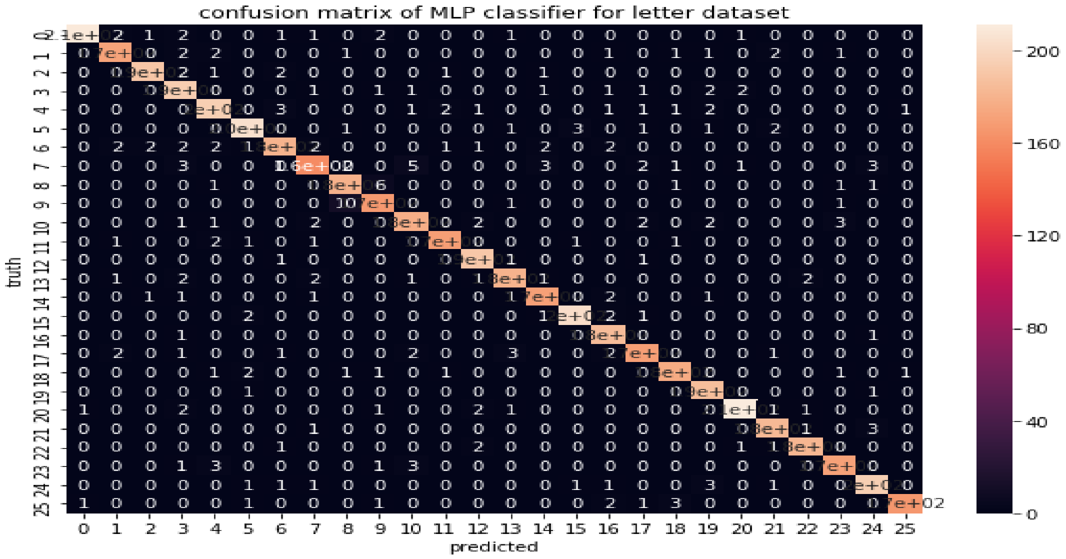


Figure 16 confusion matrix 12

4.7. Naive-Bayes

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem.

The fundamental Naive Bayes assumption is that each feature makes an:

- independent
- equal

Contribution to the outcome.

The three Naive Bayes classifiers are:

- Gaussian this naive Bayes classifier assumes that features follow a normal distribution.
- Multinomial this algorithm is used for discrete counts. For example in text-document, to count the occurrences of some words we should use Multinomial Naive-Bayes for better result.
- Bernoulli this naïve Bayes classifier is used when feature vectors are binary (means values are either 0 or 1)

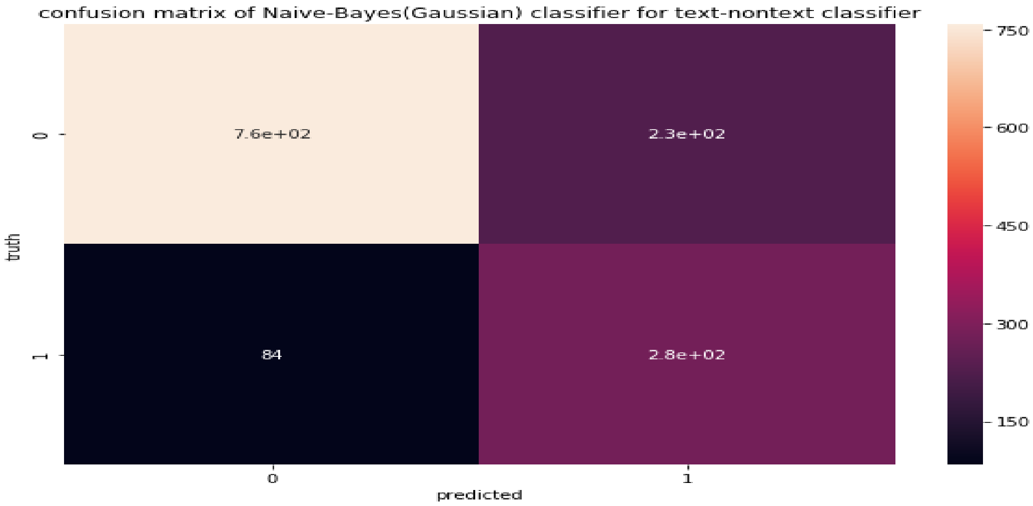


Figure 17 confusion matrix 13

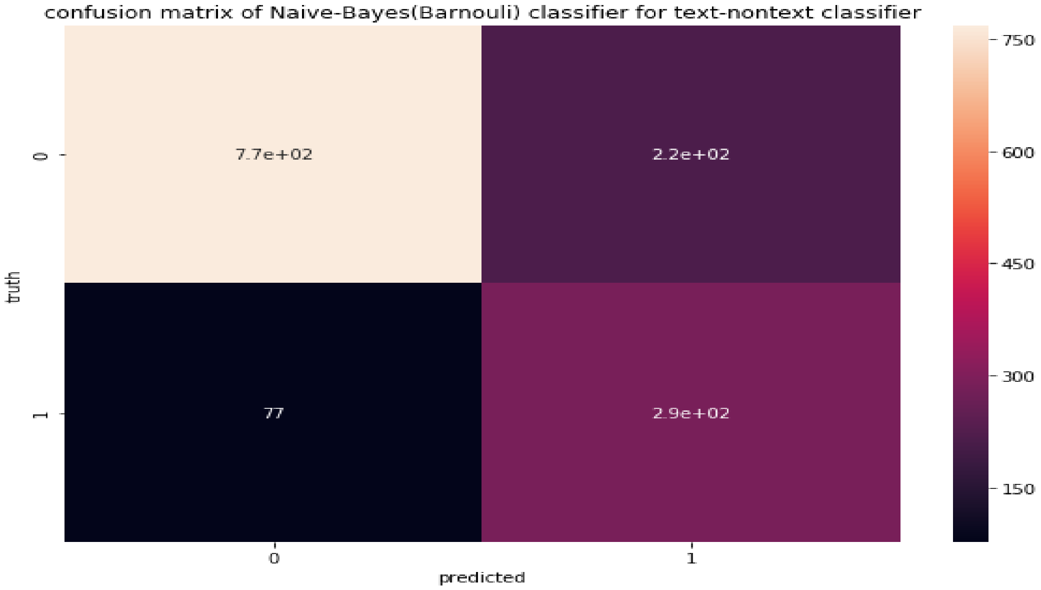


Figure 18 confusion matrix 14

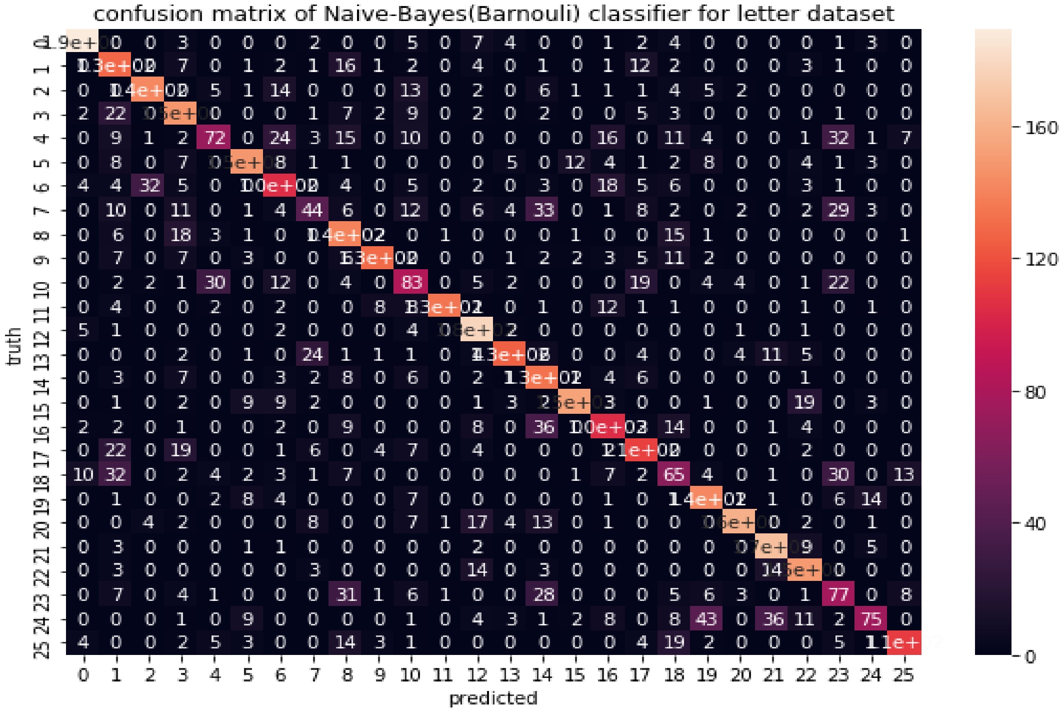


Figure 19 confusion matrix

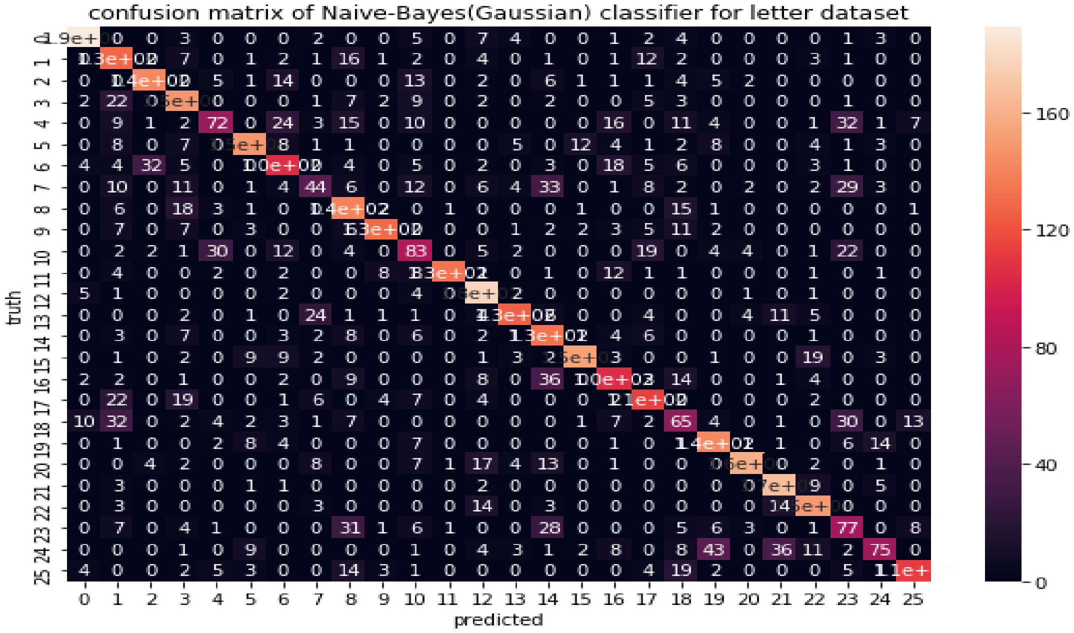


Figure 20 confusion matrix

CHAPTER FIVE

SETUP

Datasets

Throughout the whole experimental process, we have mainly focused on the standard dataset i.e. **ICDAR2015.** This dataset is text, non-text recognition dataset.

In our daily life, text detection and text recognition methodology have many applications. It can be a useful part of navigation devices when it effectively recognizes the text on the traffic street signs[9][10], It can also be an essential tool to guide the blind or the visually impaired people [11], when a URL is detected can be combined with a browser to navigate to this website. It can be very useful in the field of industrial automation. For example, when a product is detected, the user can get the information like price, or ingredients and expiration date, or specifications and size for devices, etc. It can also recognize a car number plate to detect unregister car and also minimize unnecessary car accident. It is also helpful for cross-lingual access.

Text data is ubiquitous, the main challenge with text classification is that the data is extremely high-dimensional and sparse. Not all classification methods are equally popular for text-data. For example, rule-based methods like SVM classifier tends to be more popular than other classifiers. KNN also provides good accuracy. Decision tree classifier is not good for text data.

Except for text and non-text recognition dataset (ICDAR2015), we have also used 11 different kinds of binary and multi-class datasets, which are

Table 4 various datasets

Dataset(name)	Attributes	sizes	Classes	Description
1>Breast Cancer Wisconsin	**30**	**569**	**binary**	**(diagnostic) dataset** **• classes=WDBC-Malignant,** **• WDBC-Benign (Binary).**
2>Wine recognition	**13**	**178**	**multiclass**	**Three class classification dataset**
3>The Olivetti's faces	**400**	**4096**	**multiclass**	**a set of face images**
4>The 20 newsgroups text	**450**	**18846**	**multiclass**	**Various kind of news dataset**
5>Handwritten digit	**75**	**1500**	**multiclass**	**Handwritten digit dataset**
6>The labeled Faces in the wild recognition	**450**	**13233**		**collection of JPEG pictures of famous people,**
7>Forest Cover types	**567**	**50000**	**multiclass**	**30X30m patches of forest in the US, collected for the task of predicting each patch's cover type**
8>Credit-g	**21**	**1000**	**multiclass**	**this dataset classifies people described by a set of attributes as good or bad credit risks.**
9>Phoneme	**5**	**5404**	**binary**	**the aim of this dataset is to distinguish between nasal (class 0) and oral sounds (class 1).**

10>ILPD	**10**	**583**	**binary**	**this data set contains 416 liver patient records and 167 nonliver patient records. The data set was collected from northeast of Andhra Pradesh, India. The class label divides the patients into 2 groups (liver patient or not).**
11>Letter	**17**	**20000**	**multiclass**	**the objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet.**

We have used nine classification techniques with K-fold cross-validation procedure i.e.

i. **Decision Tree** classifier (criterion=entropy, max_depth=3).
ii. **Logistic Regression** classifier.
iii. **K-Nearest-Neighbor** classifier (neighbors=15).
iv. **Random Forest** classifier (estimators=100).
v. **Multilayer Perceptron** classifier (max_iteration=1500, random_state=300).
vi. **Support Vector Machine** classifier (gamma=scale).
vii. **Bernoulli-Naive-Bayes** classifier (binarize=0.1).
viii. **Multinomial-Naive-Bayes** classifier.
ix. **Gaussian-Naive-Bayes** classifier.

CHAPTER SIX

RESULTS

A. SIGNIFICANCE ANALYSIS OF CLASSIFICATION MODELS APPLIED ON TEXT NON-TEXT DATASET

On the dataset of text non-text dataset, we have applied a few numbers of classification techniques. According to the accuracy scores of the classifier models KNN, SVM, MLP have given similar accuracy score (86%,85.5%, and 85.7% respectively). And the Random Forest classification model has given good accuracy(89%) comparatively.

All the ROC curves for different classifiers

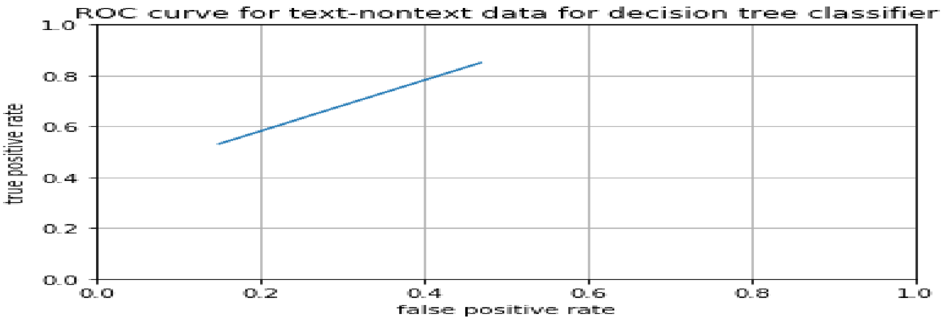


Figure 21 ROC curve 1

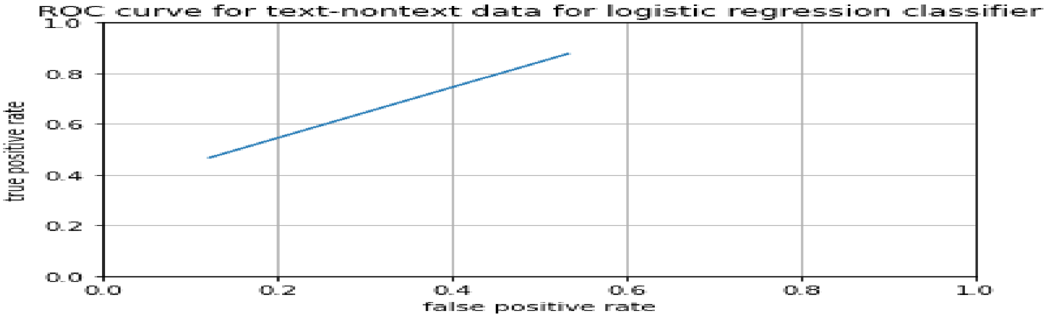


Figure 22 ROC curve 2

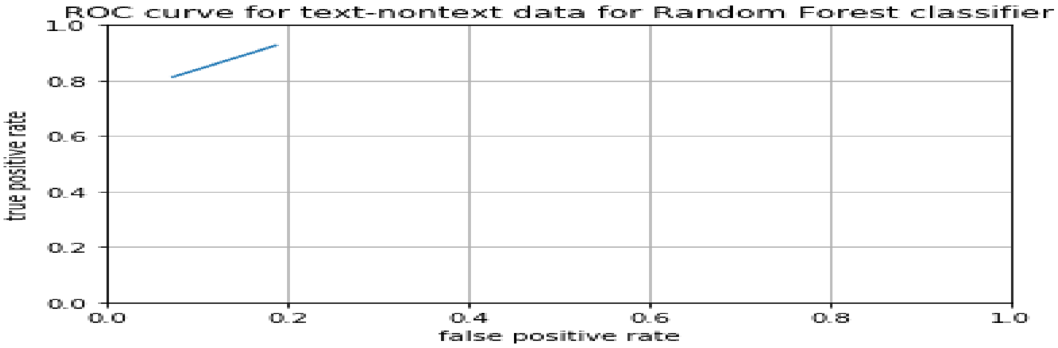


Figure 23 ROC curve 3

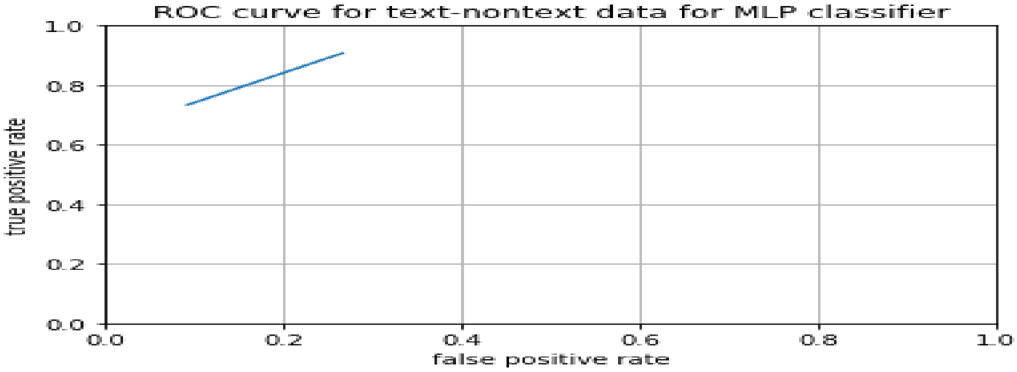


Figure 24 ROC curve 4

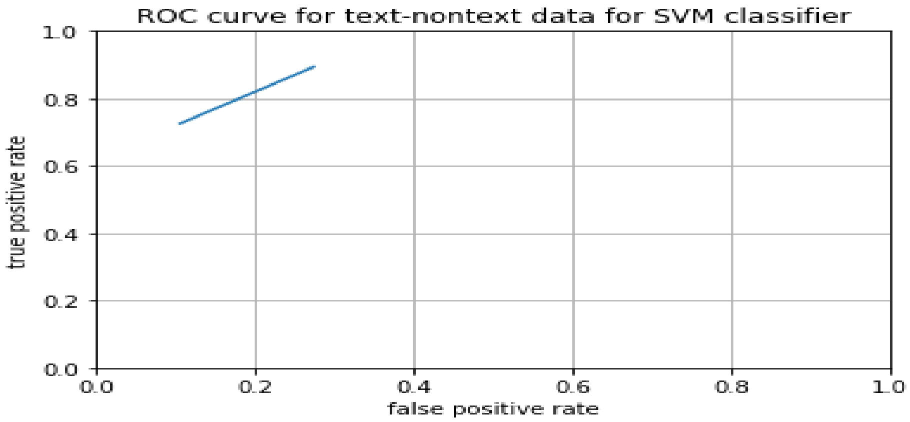


Figure 25 ROC curve 5

From the above ROC curves, we can observe that the ROC curves of KNN, SVM, and MLP look similar.

So for significance analysis, we will consider the performance metrics of those three classifier models**.**

Boxplot Diagrams of different evaluation metrics of KNN, MLP and SVM classifiers

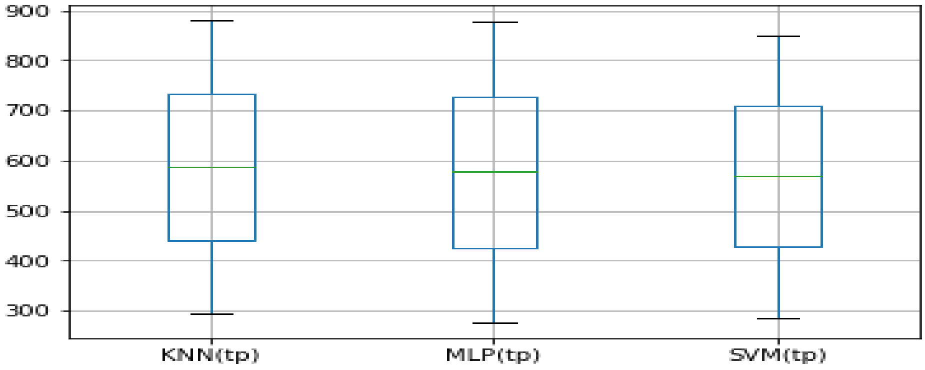


Figure 26 boxplot 1

From the boxplot diagram, we can observe that the means of true-positive values getting from KNN, MLP, and SVM classifier models are very close to each other.

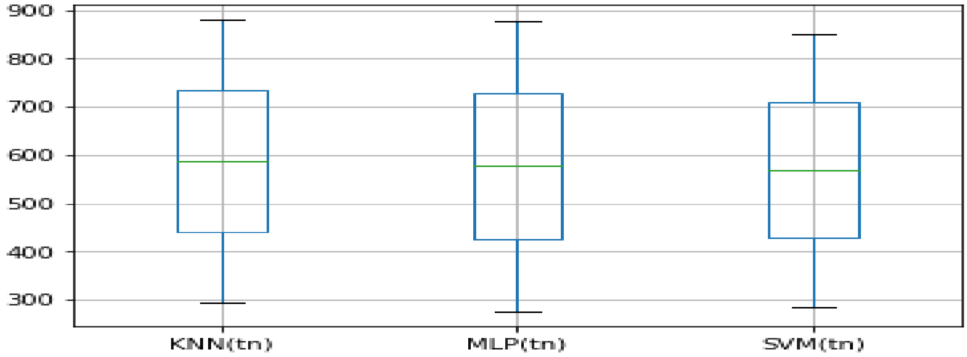


Figure 27 boxplot 2

From the boxplot diagram, we can observe that the means of true-negative values getting from KNN, MLP, and SVM classifier models are very close to each other.

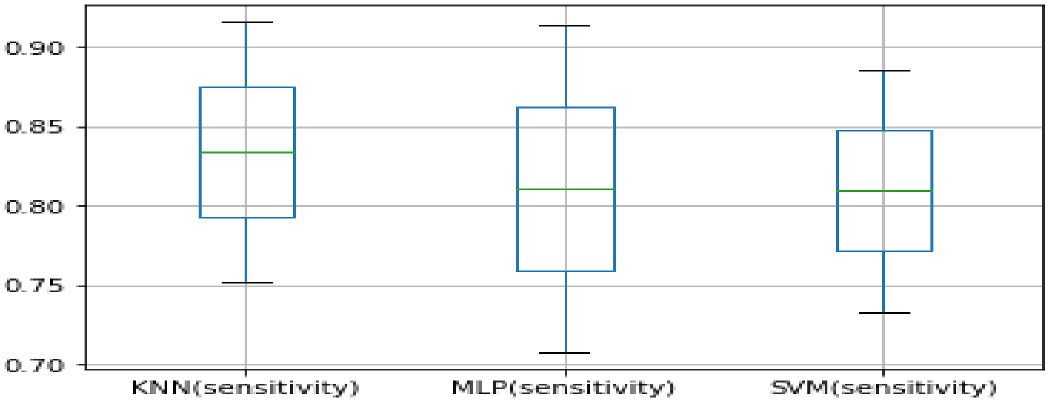


Figure 28 boxplot 3

From the boxplot diagram, we can observe that the means of sensitivity values getting from KNN, MLP, and SVM classifier models are very close to each other. All the means lie between the interval [0.80,0.85].

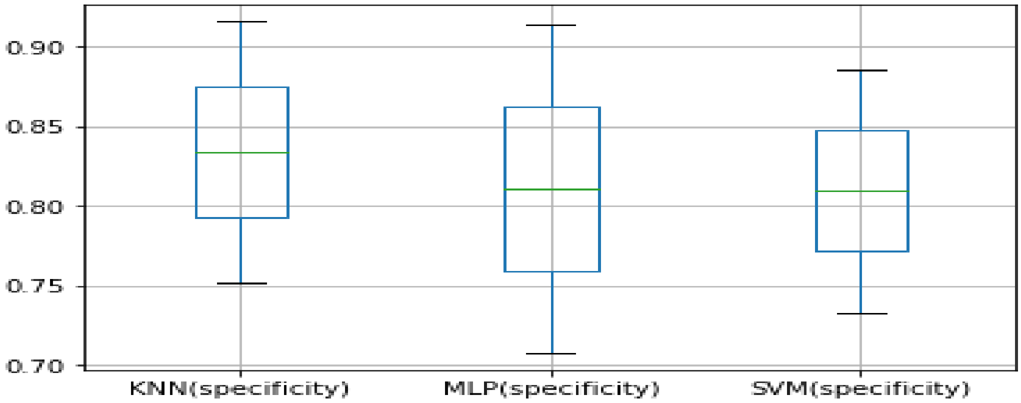


Figure 29 boxplot 4

From the boxplot diagram, we can observe that the means of specificity values getting from KNN, MLP, and SVM classifier models are very close to each other. All the means lie between the interval [0.80,0.85].

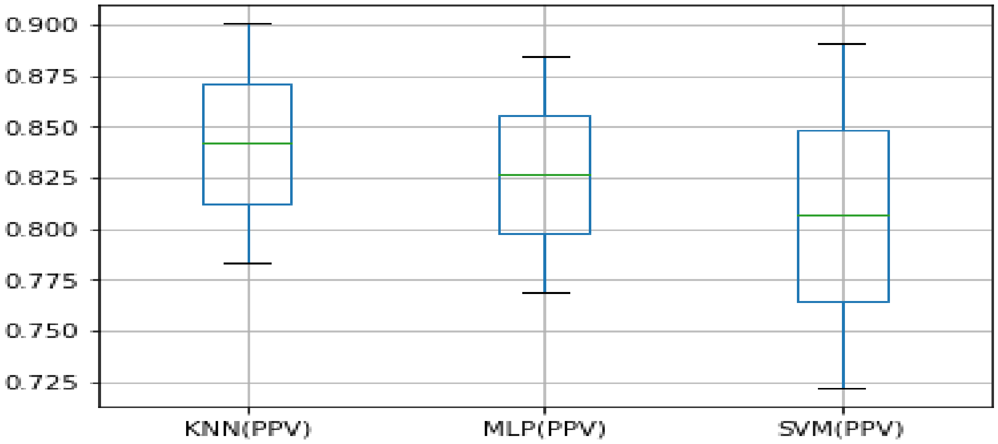


Figure 30 boxplot 5

From the boxplot diagram, we can observe that the means of PPV values getting from KNN, MLP, and SVM classifier models are very close to each other. All the means lie between the interval [0.80,0.85].

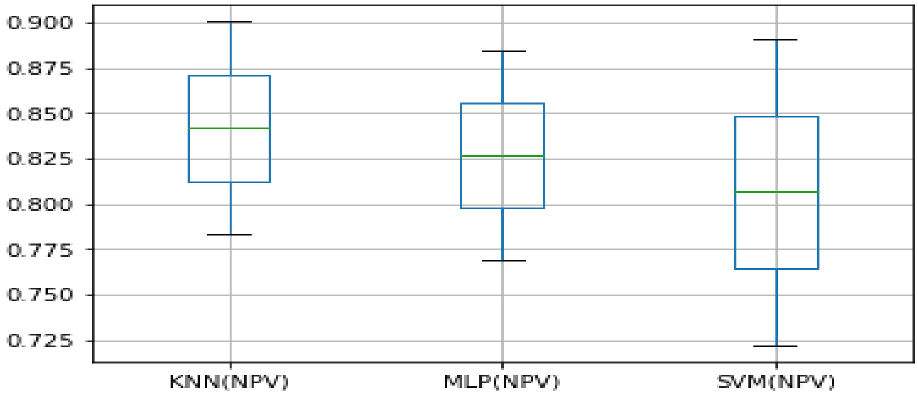


Figure 31 boxplot 6

From the boxplot diagram, we can observe that the means of NPV values getting from KNN, MLP, and SVM classifier models are very close to each other. All the means lie between the interval [0.80,0.85].

A.1. ONE SAMPLE PARAMETRIC TEST

A.1.1. T-TEST

For analysis the T-test we consider MLP and SVM classifications result applied on text non-text recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 85%).

Table 5 p-value table

True-positive	**True-negative**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
0.3556	**0.355**	**0.1708**	**0.1708**	**0.1708**	**0.1708**

We have considered a 5% confidence interval than in all cases of true-positive, true-negative, sensitivity, specificity, PPV, NPV, we accept the null hypothesis, that means the performances are the same.

A.2. TWO SAMPLES PARAMETRIC TEST

A.2.1. TWO-SAMPLE-T-TEST

For analysis the two-sample-T-test we consider MLP and SVM classifications result applied on text non-text recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 85%).

Table 6 p-value table

True-positive	**True-negative**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
0.986	**0.9862**	**0.938**	**0.938**	**0.8894**	**0.8894**

We have considered a 5% confidence interval than in all cases of true-positive, true-negative, sensitivity, specificity, PPV, NPV, we accept the null hypothesis, that means the performances are the same.

A.2.2. FISHER'S-EXACT-TEST

For analysis the Fisher's-Exact-test we consider MLP and SVM classifications result applied on text non-text recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 85%).

Contingency table:-

The contingency table shows all agreements and disagreements of the two classifiers for each testing dataset.

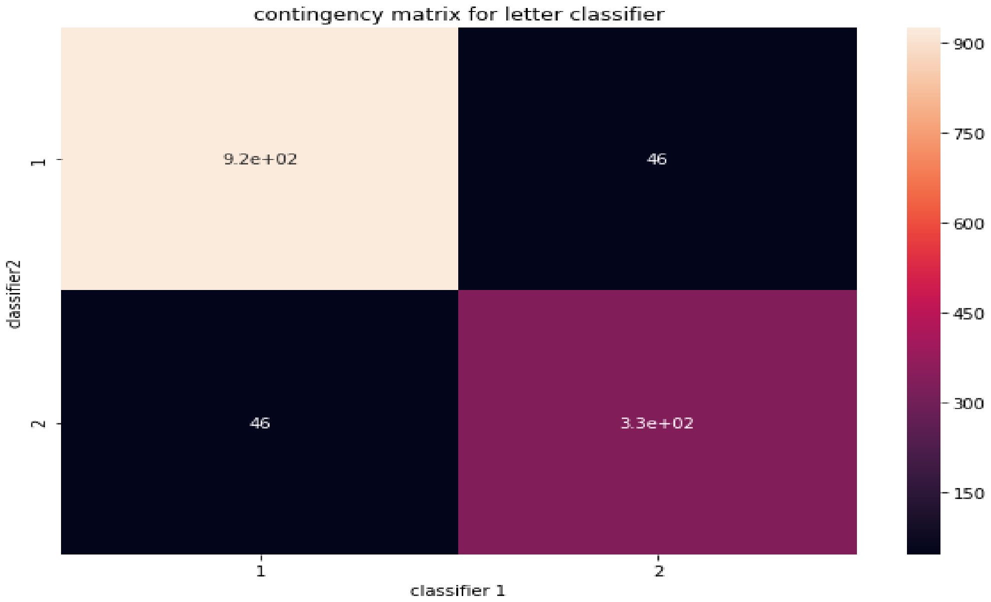


Figure 32 contingency matrix 1

Table 7 p-value table

Odd-ratios	p-value
107.53246753246754	**7.315105425010602e-193**

We have considered a 5% confidence interval. From the above table, we can see that the p-value is less than 0.05(i.e. 5%). So we reject the null hypothesis, that means the two models do not agree in the same way for every testing dataset.

A.3. Three or more samples parametric test

A.3.1. ANOVA test

For analysis the ANOVA-test we consider MLP, KNN and SVM classifications result applied on text non-text recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 85%).

Table 8 p-value table

True-positive	**True-negative**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
0.99971192	**0.9997**	**0.9966**	**0.9966**	**0.9739**	**0.9739**

We have considered a 5% confidence interval than in all cases of true-positive, true-negative, sensitivity, specificity, PPV, NPV, we accept the null hypothesis, that means the performances are the same.

A.3.2. BARTLETT'S TEST

For analysis the Bartlett's-test we consider MLP, KNN and SVM classifications result applied on text non-text recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 85%).

Table 9 p-value table

True-positive	**True-negative**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
0.99	**0.999**	**0.9786**	**0.9786**	**0.919425**	**0.919425**

We have considered a 5% confidence interval than in all cases of true-positive, true-negative, sensitivity, specificity, PPV, NPV, we accept the null hypothesis, that means the performances are the same.

A.4. NON-PARAMETRIC TEST

A.4.1 CHI-SQUARE TEST

For analysis the Chi-Square-test we consider MLP and SVM classifications result applied on text non-text recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 85%).

Observed values:-

The observed values of all agreements and disagreements for the two classifiers.

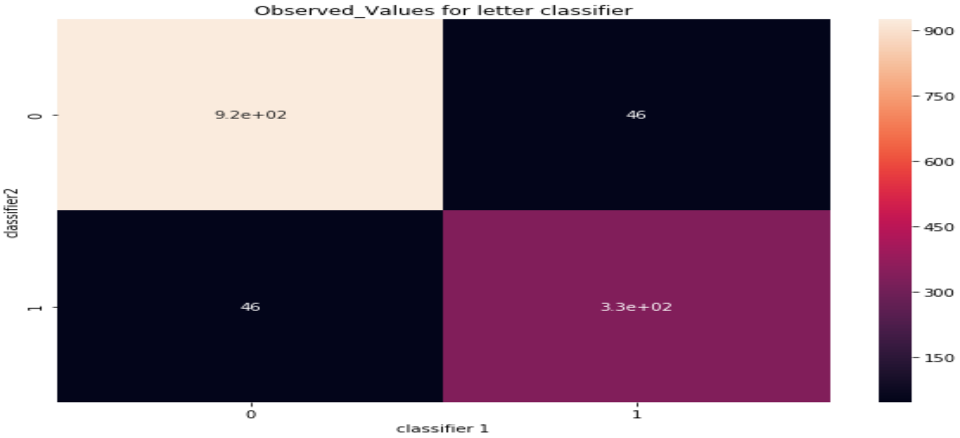


Figure 33 observed value 1

Expected values:-

The expected values of all agreements and disagreements for the two classifiers, for that we can accept the null hypothesis.

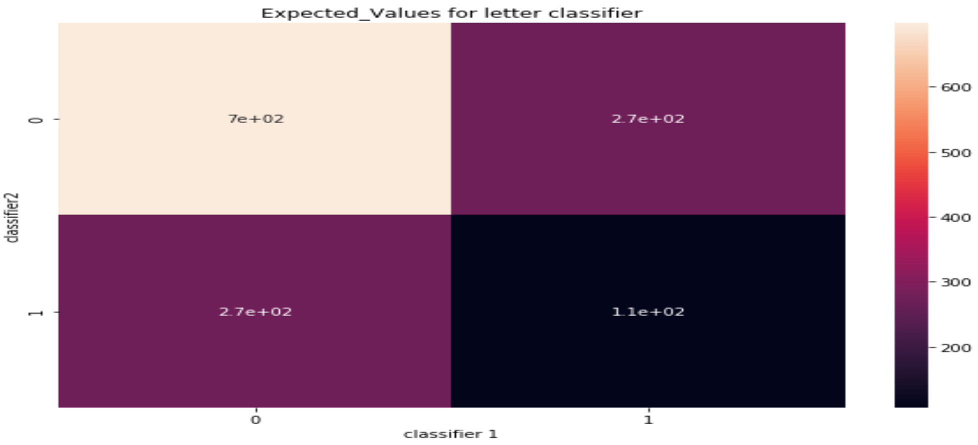


Figure 34 expected value 1

Table 10 chi-square table

Chi-square statistic	Critical value	p-value	Significance-level	Degrees of freedom
873.0044113516	3.8414588	0.0	0.05	1

We have considered a 5% confidence interval. From the above table, we can see that the p-value is less than 0.05(i.e. 5%). So we reject the null hypothesis, that means the two models do not agree in the same way for every testing data element.

A.4.2. MANN-WHITNEY TEST

For analysis the Mann-Whitney-test we consider MLP and SVM classifications result applied on text non-text recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 85%).

Table 11 p-value table

True-positive	True-negative	Sensitivity	Specificity	PPV	NPV
0.6985	0.6650055	0.665005	0.6650055	0.6650	0.6650

We have considered a 5% confidence interval than in all cases of true-positive, true-negative, sensitivity, specificity, PPV, NPV, we accept the null hypothesis, that means the performances are the same.

A.4.3. MEDIAN TEST

For analysis the Median-test we consider MLP, KNN and SVM classifications result applied on text non-text recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 85%).

Table 12 p-value table

True-positive	True-negative	Sensitivity	Specificity	PPV	NPV
1.0	1.0	0.9	0.9	0.8	0.8

We have considered a 5% confidence interval than in all cases of true-positive, true-negative, sensitivity, specificity, PPV, NPV, we accept the null hypothesis, that means the performances are the same.

A.4.4. KOLMOGOROV-SMIRNOV TEST

For analysis the Kolmogorov-Smirnov-test we consider MLP and SVM classifications result applied on text non-text recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 85%).

Table 13 p-value table

True-positive	**True-negative**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
0.8438	**1.0**	**1.0**	**0.88**	**0.8**	**0.8**

We have considered a 5% confidence interval than in all cases of true-positive, true-negative, sensitivity, specificity, PPV, NPV, we accept the null hypothesis, that means the performances are the same.

Now from all analysis, we can conclude that for text non-text dataset the classifier models KNN, SVM, MLP act more or less similar. Though the two classifier models do not agree for a single data in the same way.

B. SIGNIFICANCE ANALYSIS OF CLASSIFICATION MODELS APPLIED TO THE LETTER

On the dataset of letter dataset, we have applied a few numbers of classification techniques. According to the accuracy scores of the classifier model, KNN and MLP have given similar accuracy score (92% and 93%respectively). And the Random Forest classification model has given good accuracy (96%) comparatively.

All the ROC curves for different classifiers

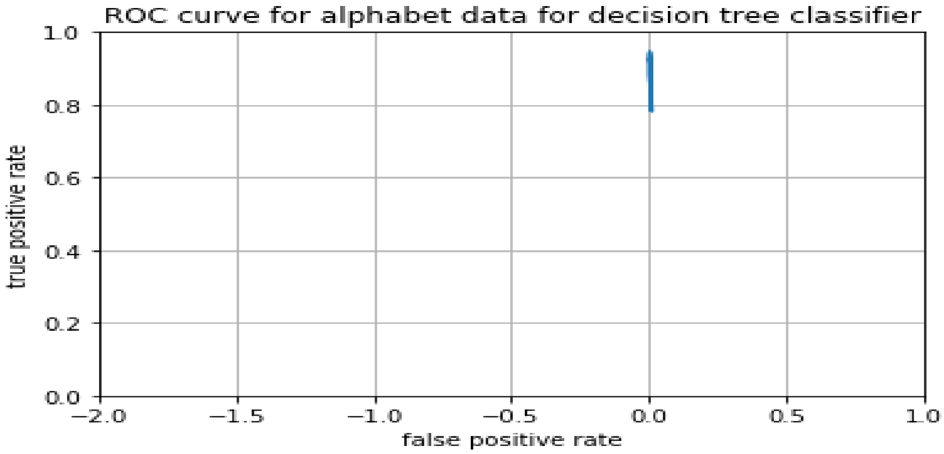


Figure 35 ROC curve 7

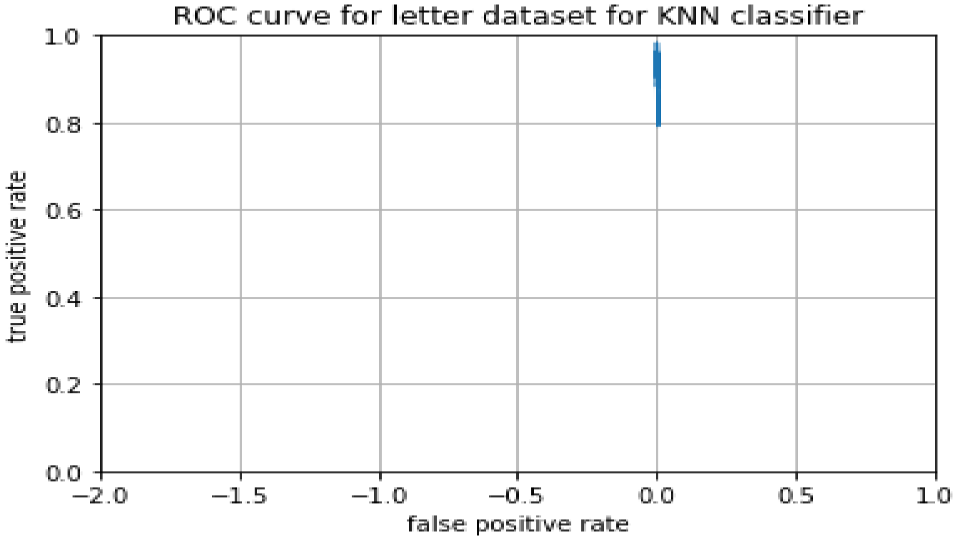


Figure 36 ROC curve 8

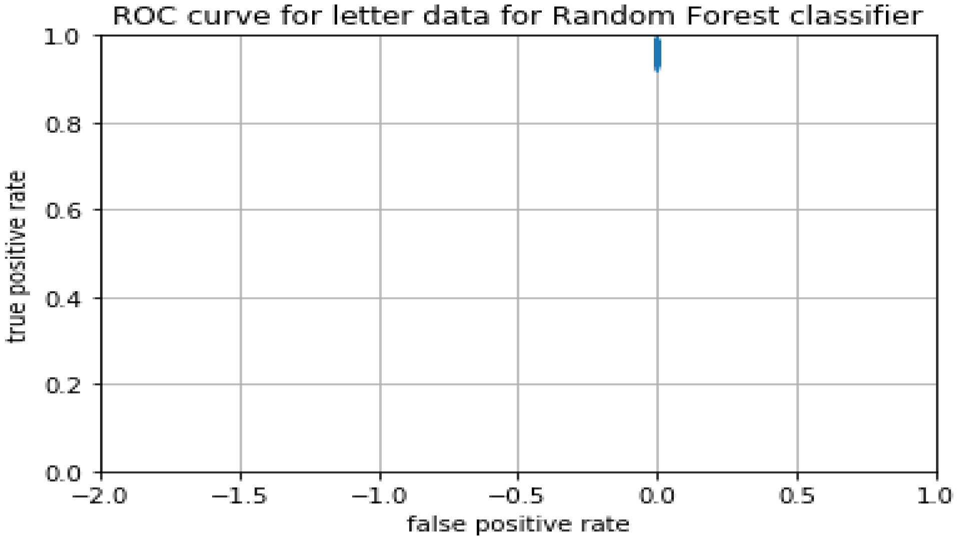


Figure 37 ROC curve 9

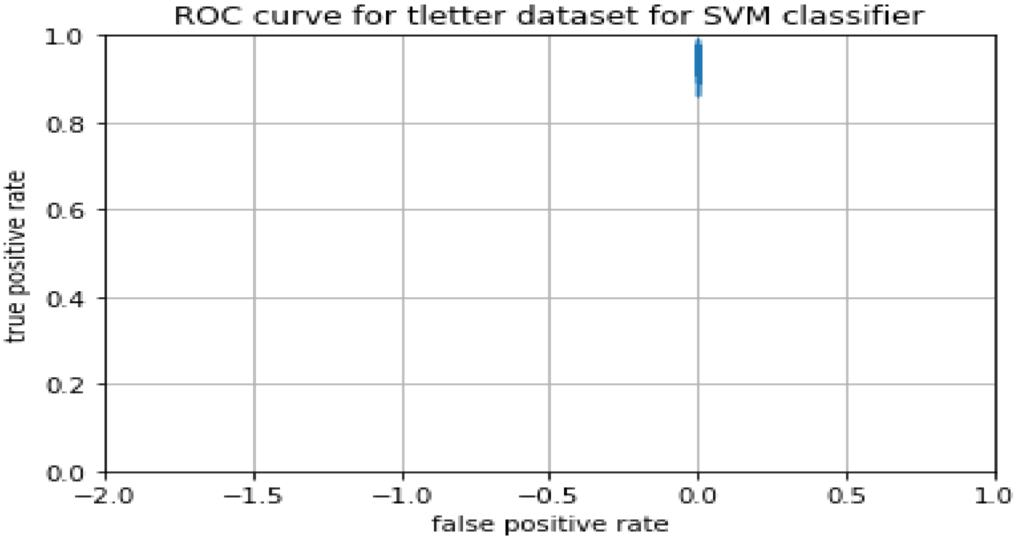


Figure 38 ROC curve 10

From the above ROC curves, we can observe that the ROC curves of SVM and MLP look similar. So for significance analysis, we will consider the performance metrics of those three classifier models

<u>Boxplot Diagrams of different evaluation metrics of MLP and SVM classifiers</u>

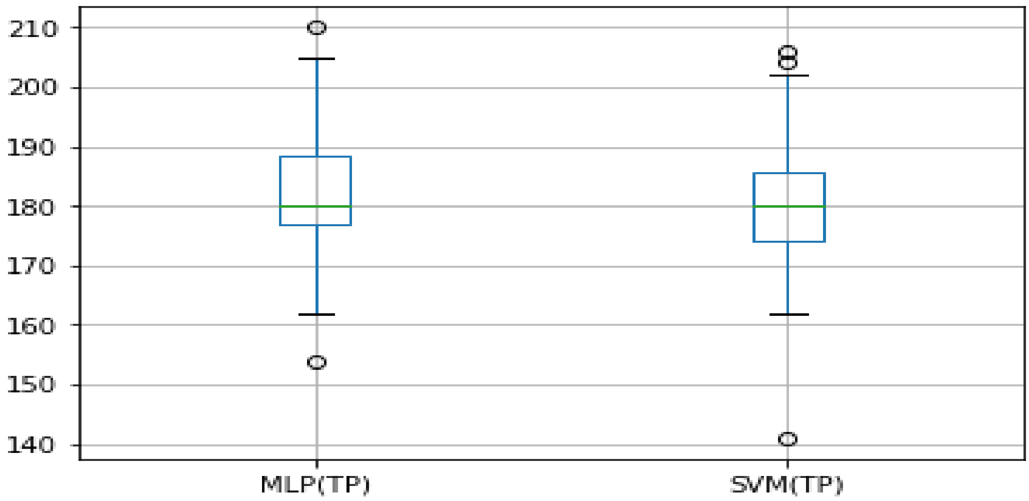


Figure 39 boxplot 9

From the boxplot diagram, we can observe that the means of true-positive values getting from MLP, and SVM classifier models are the same.

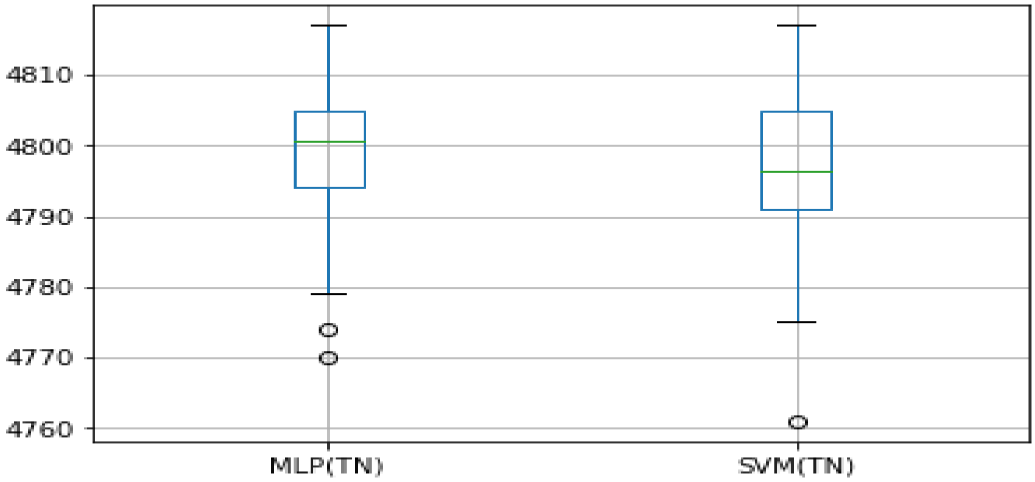


Figure 40 boxplot 10

From the boxplot diagram, we can observe that the means of true-negative values getting from MLP, and SVM classifier models are not so close to each other.

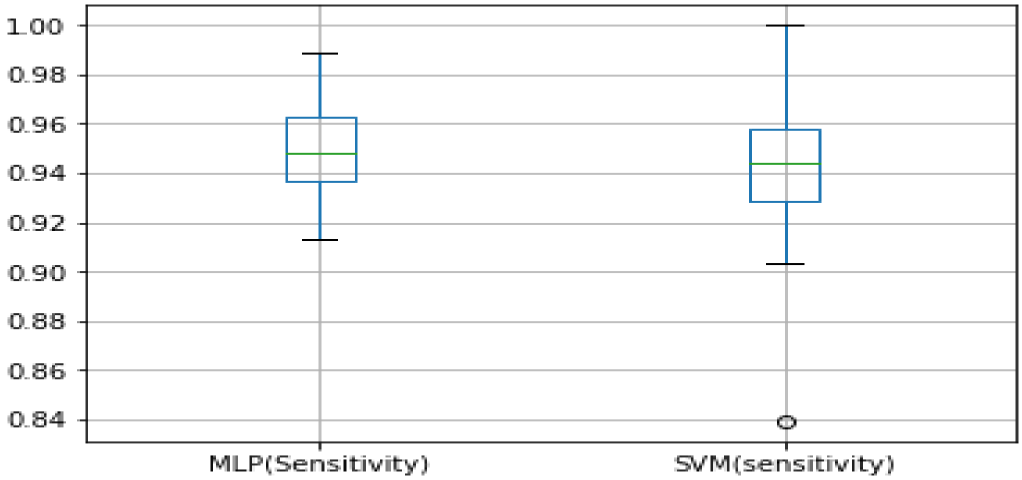


Figure 41 boxplot 11

From the boxplot diagram, we can observe that the means of sensitivity values getting from MLP, and SVM classifier models are very close to each other. All the means lie between the interval [0.94,0.96].

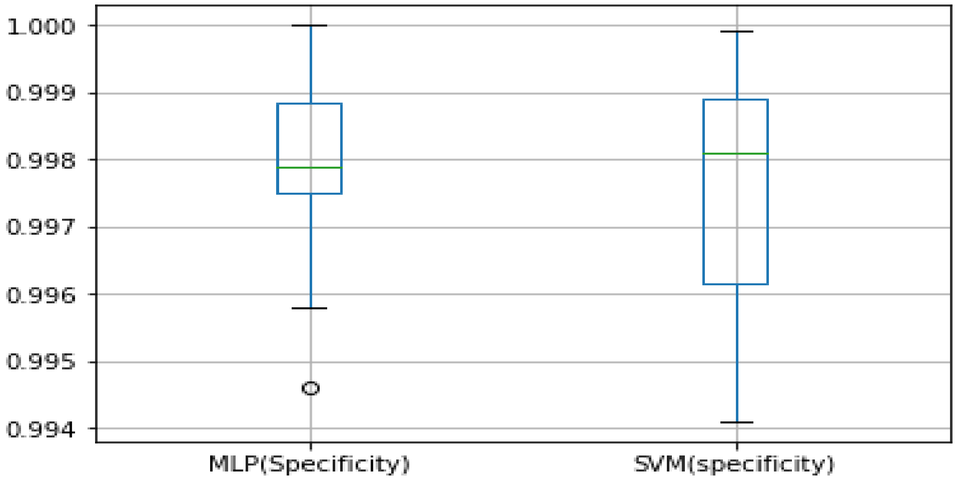


Figure 42 boxplot 12

From the boxplot diagram, we can observe that the means of specificity values getting from MLP, and SVM classifier models are very close to each other. All the means lie between the interval [0.997,0.999].

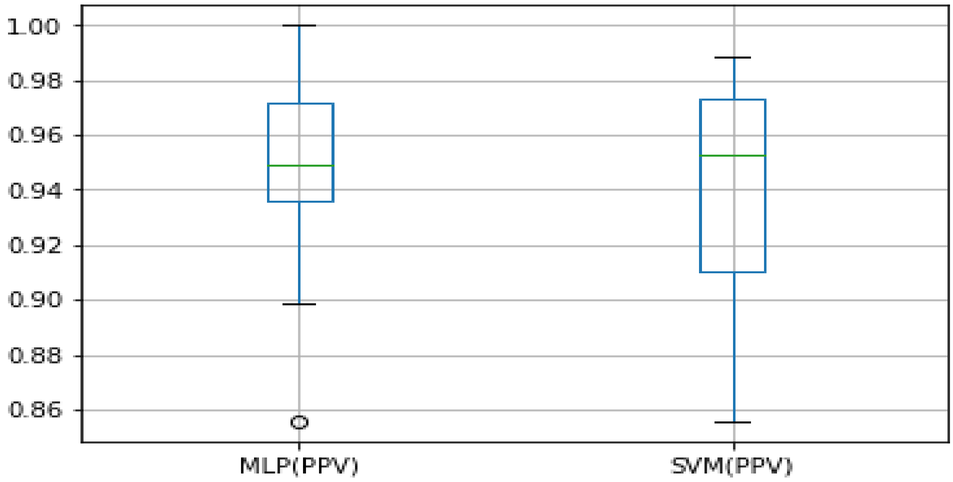


Figure 43 boxplot 13

From the boxplot diagram, we can observe that the means of PPV values getting from MLP, and SVM classifier models are very close to each other. All the means lie between the interval [0.94,0.96].

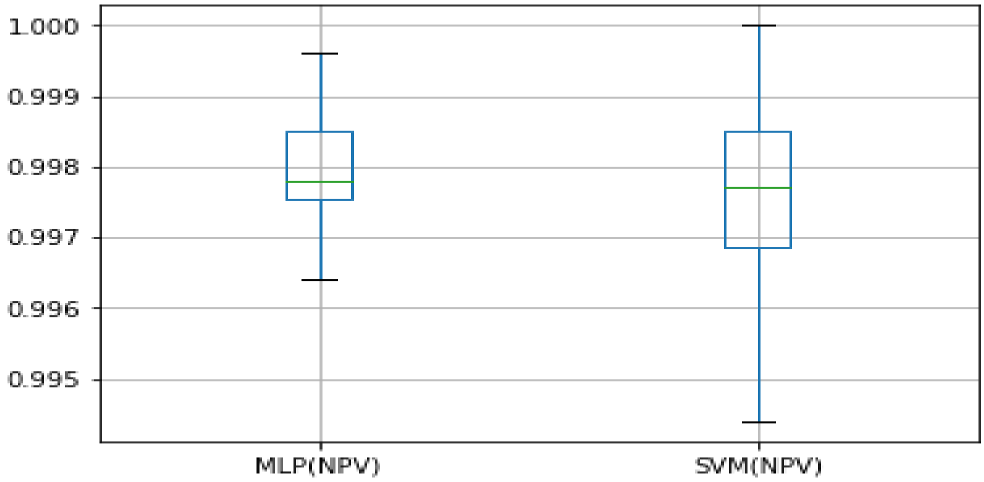


Figure 44 boxplot 14

From the boxplot diagram, we can observe that the means of NPV values getting from MLP, and SVM classifier models are very close to each other. All the means lie between the interval [0.997,0.998].

B.1. ONE SAMPLE PARAMETRIC TEST

B.1.1. T-TEST

For analysis the T-test we consider MLP and SVM classifications result applied on letter recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 94%).

Table 14 p-value table

True-positive	**True-negative**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
0.874334	**0.4999**	**0.9004**	**0.9004**	**0.9004**	**0.9004**

We have considered a 5% confidence interval than in all cases of true-positive, true-negative, sensitivity, specificity, PPV, NPV, we accept the null hypothesis, that means the performances are the same.

B.2. TWO SAMPLES PARAMETRIC TEST

B.2.1. TWO-SAMPLE-T-TEST

For analysis the two-sample-T-test we consider MLP and SVM classifications result applied on letter recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 94%).

Table 15 p-value table

True-positive	True-negative	Sensitivity	Specificity	PPV	NPV
0.6175	**0.57749**	**0.1491**	**0.3218**	**0.35206**	**0.1366**

We have considered a 5% confidence interval than in all cases of true-positive, true-negative, sensitivity, specificity, PPV, NPV, we accept the null hypothesis, that means the performances are the same.

B.3. NON-PARAMETRIC TEST

B.3.1 CHI-SQUARE TEST

For analysis the Chi-Square-test we consider MLP and SVM classifications result applied on text non-text recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 94%).

Contingency Matrix

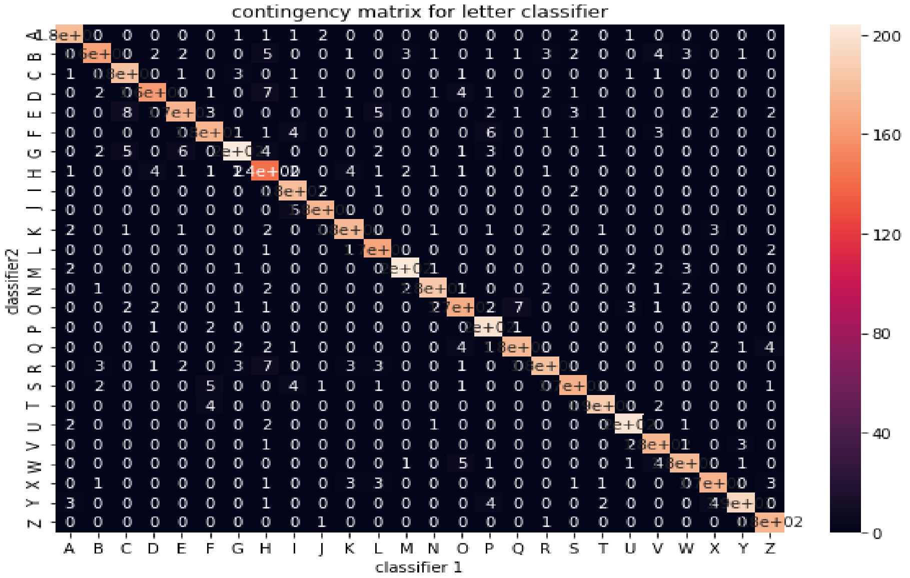


Figure 45 contingency matrix 2

Observed values:-

The observed values of all agreements and disagreements for the two classifiers.

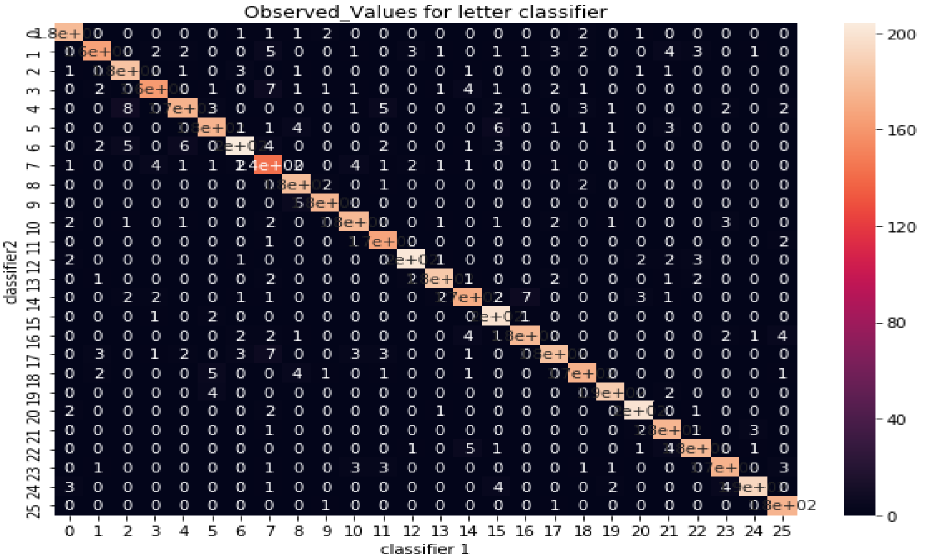


Figure 46 observed value 2

Expected values:-

The expected values of all agreements and disagreements for the two classifiers, for that we can accept the null hypothesis.

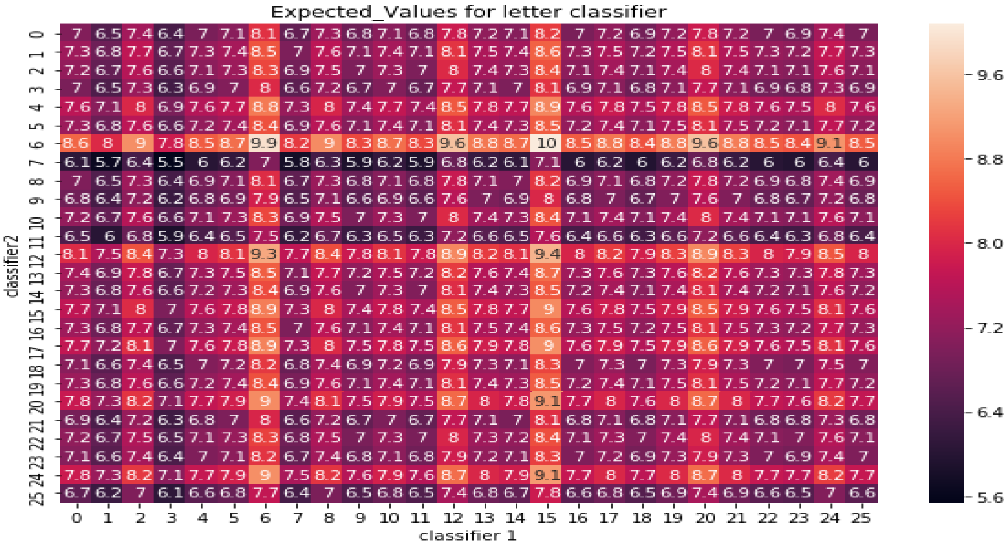


Figure 47 expected value 2

Table 16 chi-square table 2

Chi-square statistic	Critical value	p-value	Significance-level	Degrees of freedom
8134.5367	3.841458	0.02	0.05	1

We have considered a 5% confidence interval. From the above table, we can see that the p-value is less than 0.05(i.e. 5%). So we reject the null hypothesis, that means the two models do not agree in the same way for every testing data element.

B.3.2. MANN-WHITNEY TEST

For analysis the Mann-Whitney-test we consider MLP and SVM classifications result applied on letter recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 94%).

Table 17 p-value table

True-positive	True-negative	Sensitivity	Specificity	PPV	NPV
0.5516	**0.79038**	**0.332**	**0.8332**	**0.9489**	**0.3275**

We have considered a 5% confidence interval than in all cases of true-positive, true-negative, sensitivity, specificity, PPV, NPV, we accept the null hypothesis, that means the performances are the same.

B.3.3. KOLMOGOROV-SMIRNOV TEST

For analysis the Kolmogorov-Smirnov-test we consider MLP and SVM classifications result applied on letter recognition dataset. We have considered the two classifiers as their accuracy scores are very close to each other (both accuracy scores are nearly 94%).

Table 18 p-value table

True-positive	True-negative	Sensitivity	Specificity	PPV	NPV
0.892	**0.63**	**0.2581**	**0.440**	**0.4401**	**0.6738**

We have considered a 5% confidence interval than in all cases of true-positive, true-negative, sensitivity, specificity, PPV, NPV, we accept the null hypothesis, that means the performances are the same.

Now from all analysis, we can conclude that for letter dataset the classifier models SVM and MLP act more or less similar. Though the two classifier models do not agree for a single data in the same way.

CHAPTER SEVEN

CONCLUSION

In this paper, we propose to use various kinds of parametric as well as non-parametric hypothetical tests to analyze the performances of classifier models on several kinds of datasets. Hypothesis testing is a practical approach to significant analysis. Without proper significant analysis, we cannot make a conclusion about the performance of any classifier model on a particular kind of dataset. For Hypothetical testing, we have used various evaluation metrics such as true-positive, true-negative, sensitivity, specificity, PPV and NPV. We know that sometimes parametric tests can mislead us to make a conclusion, as all the parametric tests make so many assumptions about the population from which the samples belong to. To avoid the mislead we also apply a number of non-parametric tests to make our conclusion stronger, as the non-parametric tests are robust in nature. Though non-parametric test gives approximate p-value this kind of tests does not mislead us. This is an interesting direction for research purpose, in future we can do significance analysis of other evaluation metrics of classifiers. This significant analysis will help us in case of semi-supervised learning where the testing datasets are not labeled. We can apply the best model for a particular dataset for semi-supervised learning with the help of significant analysis of all classifiers in case of supervised learning. We have only considered classification problems, we can try a significance analysis of regression problems.

REFERENCES

[1] D. P. Classification, “How do we compare the relative performance among competing models? 1.”

[2] G. Bontempi and S. Ben Taieb, “Statistical foundations of machine learning,” *OTexts,* pp. 1–45, 2015.

[3] S. McKillup and S. McKillup, “Non-parametric statistics,” in *Statistics Explained*, 2012.

[4] O. Taner, O. Aslan, and E. Alpayd, “Multivariate Statistical Tests for Comparing,” pp. 1–15, 2011.

[5] K. R. CLARKE, “Non-parametric multivariate analyses of changes in community structure,” *Aust. J. Ecol.*, 1993.

[6] “CHAPTER 6 Hypothesis Testing,” in *Statistics for Marketing and Consumer Research*, 2011.

[7] V. Grech and N. Calleja, “WASP (Write a Scientific Paper): Parametric vs. non-parametric tests,” *Early Hum. Dev.*, 2018.

[8] P. Sedgwick, “A comparison of parametric and non-parametric statistical tests,” *BMJ (Online)*. 2015.

[9] J. Yang, J. Gao, Y. Zhang, X. Chen, and A. Waibel, “An Automatic Sign Recognition and Translation System,” in *Proceedings of the 2001 Workshop on Perceptive User Interfaces*, 2001, pp. 1–8.

[10] A. González, L. M. Bergasa, J. J. Yebes, and J. Almazán, “Traffic panels detection using visual appearance,” in *2013 IEEE Intelligent Vehicles Symposium (IV)*, 2013, pp. 1221–1226.

[11] A. Zandifar and A. Chahine, “A video-based interface to textual information for the visually impaired,” in *Proceedings of the 4th IEEE International Conference on Multimodal interfaces*, 2002, p. 325.