IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Comparative Study of Different Classification Algorithms for Early Prediction of Cancer

Shekhar Tanwar[1], **Shalini L.**[2]

Alumni, School of Computing Science and Engineering, VIT, Vellore, India[1]

Assistant Professor (Senior), School of Computing Science and Engineering, VIT, Vellore, India[2]

Abstract: Breast Cancer, one of the most common diseases which has impacted the female population is a result of two genes BRCA1 and BRCA2. The geneses result in the formation of cysts or lumps in the female breast which can later develop into a fully developed tumor. The tumor can either be malignant (cancerous) or benign (harmless), depending on the composition of the nuclei which forms it. This case study focuses on the several characteristics of the lumps and using classification algorithms makes an attempt for early prediction of cancer symptoms depending on the various characteristics of the lump.

Keywords: Re-index, Correlation Analysis, Relativity Analysis, 10-fold Cross Validation, Logistic Regression, Naïve Bayes, Gradient Boosted Trees, Random Forest Trees, ROC Curves, Precision Recall Curves.

I. INTRODUCTION

The main focus of this paper is to study the impact of different classification algorithms in the prediction of label attributes. The model will be judged using Accuracy, Precision and Recall and ROC curves as parameters. These parameters come in handy when the model is first trained on Train data and then on Test Data. This paper is catalogued as follows, Section II. presents a related work in this field. Section III. discusses the methodology and the aspects of classification algorithms and respective datasets. Section IV. Elaborates Experiment and finalizes the results produced by the algorithms. Section V. presents the detailed conclusion.

II. RELATED WORK

Dr. Wolberg and Prof. Mangasarian along with his two students focused on Fine Needle Aspiration (FNA) to accurately diagnose breast masses. Out of all the characteristics of the FNA sample, Certain characteristics proved to more significant in contributing towards diagnosis. The team constructed a classifier called the multisurface method (MSM) and using the most significant features accurately recognized 97% of the new cases.

The recognition process was constructed using the following process:

- A FNA sample was taken from the breast mass, and the individual well differentiated cellular nuclei were identified.
- The individual nuclei were isolated, and the classification algorithm used computed the mean, standard error and extreme values resulting on total of 30 nuclear features for each sample.
- Based on 569 cases, a linear classifier was constructed to differentiate benign from malignant samples.

The system has been successful so far, and to this date has correctly diagnosed 176 consecutive new patients (119 benign, 57 malignant)

III. METHODOLOGY

This case study makes use of the following procedures to for the classification problem:

- A dataset is selected which comprises of values of various features present in the cells of breast tissue
- To optimize the application of classification algorithms, the dataset re-indexed using the randomly permutation function.
- A correlation analysis is applied between the characteristics to compute the correlation
- To group characteristics of each type, relativity analysis is performed on the dataset
- The dataset is partitioned in the ratio 7:3 between Train and Test Dataset
- To optimize and accurately classify the train dataset into Malignant and Benign, 10-fold Cross Validation is performed on the dataset.

DOI10.17148/IJARCCE.2017.6844

- The classification algorithms, Logistic Regression, Naïve Bayes, Gradient Boosted Trees, Random Forest Trees are applied on the Train dataset and, and their corresponding accuracy on the Test dataset is judged using ROC Curves and Precision Recall Curves

A. Dataset Used

The dataset used is provided by University of Wisconsin Madison, and comprises of values for several features present in the breast tissue for each cellular nuclei in the tissue. The dataset comprises of 699 rows along with 11 columns and is first reshuffled and then all the missing values in the dataset are replaced with 0 for accuracy of results.

B. Classifier Used

1) Logistic Regression

Logistic Regression is a type of regression analysis, which is used when the dependent variable is dichotomous(binary). Similar to other regression analysis, Logistic Regression is a kind of prediction algorithm, which is used to describe a relation between one dependent variable and other one or more independent variable(s).
Logistic Regression works on the following assumptions:

- The outcome must be discrete that is to say that the dependent variable should be dichotomous in nature
- The dataset should not comprise of any outliers and even there are such values, they need to either standardized or only a range of z scores need to be collected
- The various features needed for classification or the assumed independent variables in this case should not have high correlations among them.

Logistic Regression estimates the log odds of an event, which can be mathematically expressed as:

$$\mathbf{g(F(x)) = \ln(\frac{F(x)}{1-F(x)}) = \beta_0 + \beta_1 x_1 + \ldots\ldots.. + \beta_n x_n}$$

Optimum efficiency of the model results when just the right number of features are used for training the model, for a dataset having too many features, training the model may result into overfitting.

2) Naïve Bayes

The Naïve Bayes classifier falls under the family of probabilistic classification algorithms and is not a single algorithm to train the model for classification, but is rather a collection of algorithms which classify documents in one category or another (i.e. legitimate text or spam, sports or politics, in this case study (malignant or benign)). These collective algorithms work on the assumption that the features being used for classification are independent of one another. These algorithms assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. Despite its simplicity, the Naïve Bayes classification algorithm can outperform more sophisticated classification methods. Naïve Bayes algorithm works on the foundation of Bayes theorem and make use of prior probability, likelihood and evidence of occurring of an event.
For a problem instance having n features (x_1,……x_n, assumed to be independent), the probability of instance variables is given by
p (C_k | x_1 ,…..,x_n), where C_k represent the possible classes in which the various instances would be classified.
Using Bayes Theorem, the conditional problem can be decomposed as:

$$\mathbf{p\ (C_k\,|x) = \frac{p(C(k))p(x|C(k))}{p(x)}}$$

$$\mathbf{p\ (C_k\,|x) = \frac{1}{Z}\,p(C_k)\pi\ (_{i\,=\,1\ to\ n})\ p\ (x_i\,|C_k)}$$

where Z = p(x), is a scaling factor dependent on x_1,x_2,…….x_n. Its value remains constant if the values of the feature variables are known.

3) Gradient Boosted Trees

Gradient Boosted Trees is a one type of classification algorithms and uses several weak models (like decision trees) and fine-tunes the models by improving the arbitrary differential loss function with each iteration. The GBT method teaches a model F to predict an output in the form y^ = F(x), and reduces the mean squared error $(y^\wedge - y)^2$, using all the actual values of the output variable present in the Train dataset.

In simpler words, at each stage the GBT method improves on some imperfect model F_m, by building a new model F_{m+1} based on it and some estimator h. It is this h, which the GBT tries to improve in its iterations 1≤m≤M, along with the means squared error. Mathematically,

$$\mathbf{F_{m+1}(x) = F_m(x) + h}$$

4) Random Forest Trees
Random forest is a collection of decision trees. It is presented independently with some controlled modification. Trees and the results included in Random Forest are based on majority of accurate output. Random forest is the best classifier for large datasets. 1) If 'n' is the number of cases in the training set, then 'n' cases are to be sampled randomly but with replacement, from the original data. This sample will act as a training set for growing the tree. 2) If input variables are 'M' in number, a number mM is specified such that at each node, m variables randomly selected out of the 'M' input variables and among all these 'm', the best split is used to split the node. The value of m is kept constant during the forest growing. 3) Each tree is made to grow to the largest extent possible. Pruning is restricted just to get more accuracy compromising increased execution time

C. Other Techniques applied for making dataset feasible for application of Classifiers
1) Reshuffling & Dealing with Missing Values
For maximum efficiency of the classification algorithms, the relative position of the features and their corresponding values are shuffled. The dataset is then re-indexed and finally the missing values in the dataset are substituted with zero, for the classifiers to understand and act on it.

2) Correlation Analysis
As one of the classifiers used is Logistic Regression, it is imperative that the values of the features used for classification are checked for their correlation with each other. All those features which have correlation below 0.9 are assumed to be fit for application of Logistic Regression.

3) Relativity Analysis
For each characteristic, excluding the index, the Relativity analysis gives the Average Response and the Frequency Distribution of other columns. This practice, offers a deeper insight into how the data is linked with each other and facilitates in picking the best characteristics for prediction.

IV. EXPERIMENT AND RESULTS

The data set is partitioned into Train and Test in the ratio 7:3 while keeping the relative percentages of the features intact in each subset. After reshuffling and re-indexing the Train dataset, the correlation between the features is computed and those features which are comparatively independent of each other are considered for model building. To check the how much each feature weighs in comparison to others, the Relativity Analysis is performed on the Train dataset and finally the 10-fold cross validation is performed for efficient and bias free classification. The various classification algorithms i.e. Logistic Regression, Naïve Bayes, Gradient Boosted Trees and Random Forest Trees are applied first on the Train dataset and then their efficiency of classification is checked on the Test Dataset. The comparative performance of these algorithms is checked using the ROC curve and the Precision Recall Curve. The adjoining tables summarize the results of the experiments.

TABLE 1. MODELS AND ACCURACIES ON TRAIN DATASET

Model Used	Accuracy
Logistic Regression	93.91%
Naïve Bayes	91.23%
Gradient Boosted Trees	89.12%
Random Forest Trees	97.42%

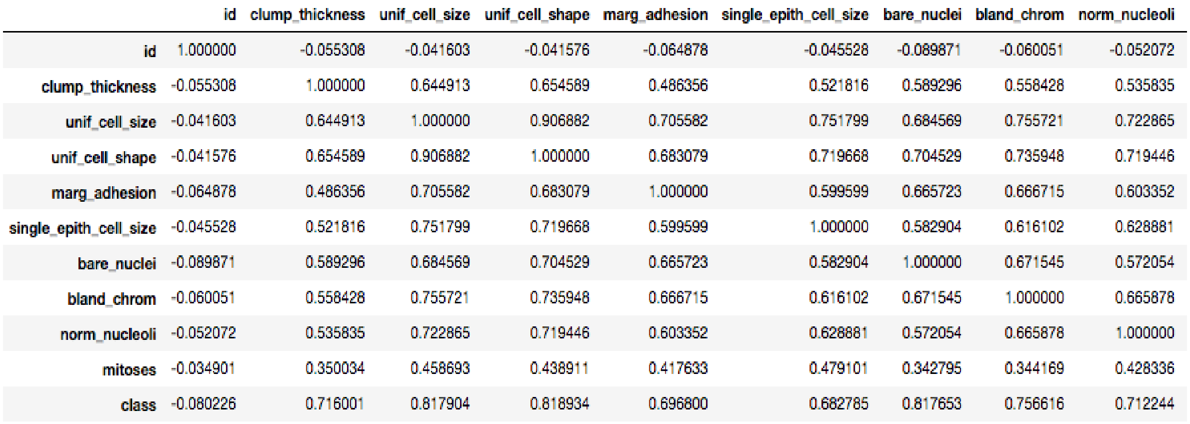
TABLE1. shows the collective display of accuracies of several classification algorithms on Train dataset. The table clearly shows that out of the several algorithms used, Random Forest, with its highest accuracy rate of classification appears to be the clear choice.

TABLE 2. MODELS AND ACCURACIES ON TEST DATASET

Model Used	Accuracy
Logistic Regression	91.90%
Naïve Bayes	86.54%
Gradient Boosted Trees	82.78%
Random Forest Trees	94.56%

DOI10.17148/IJARCCE.2017.6844

TABLE 2. shows the collective display of accuracies of several classification algorithms on Test dataset. It becomes evident that Random Forest has performed reasonably well among the classification algorithms used.



	id	clump_thickness	unif_cell_size	unif_cell_shape	marg_adhesion	single_epith_cell_size	bare_nuclei	bland_chrom	norm_nucleoli
id	1.000000	-0.055308	-0.041603	-0.041576	-0.064878	-0.045528	-0.089871	-0.060051	-0.052072
clump_thickness	-0.055308	1.000000	0.644913	0.654589	0.486356	0.521816	0.589296	0.558428	0.535835
unif_cell_size	-0.041603	0.644913	1.000000	0.906882	0.705582	0.751799	0.684569	0.755721	0.722865
unif_cell_shape	-0.041576	0.654589	0.906882	1.000000	0.683079	0.719668	0.704529	0.735948	0.719446
marg_adhesion	-0.064878	0.486356	0.705582	0.683079	1.000000	0.599599	0.665723	0.666715	0.603352
single_epith_cell_size	-0.045528	0.521816	0.751799	0.719668	0.599599	1.000000	0.582904	0.616102	0.628881
bare_nuclei	-0.089871	0.589296	0.684569	0.704529	0.665723	0.582904	1.000000	0.671545	0.572054
bland_chrom	-0.060051	0.558428	0.755721	0.735948	0.666715	0.616102	0.671545	1.000000	0.665878
norm_nucleoli	-0.052072	0.535835	0.722865	0.719446	0.603352	0.628881	0.572054	0.665878	1.000000
mitoses	-0.034901	0.350034	0.458693	0.438911	0.417633	0.479101	0.342795	0.344169	0.428336
class	-0.080226	0.716001	0.817904	0.818934	0.696800	0.682785	0.817653	0.756616	0.712244

Fig3. Table Summarizing Correlation between the characteristics

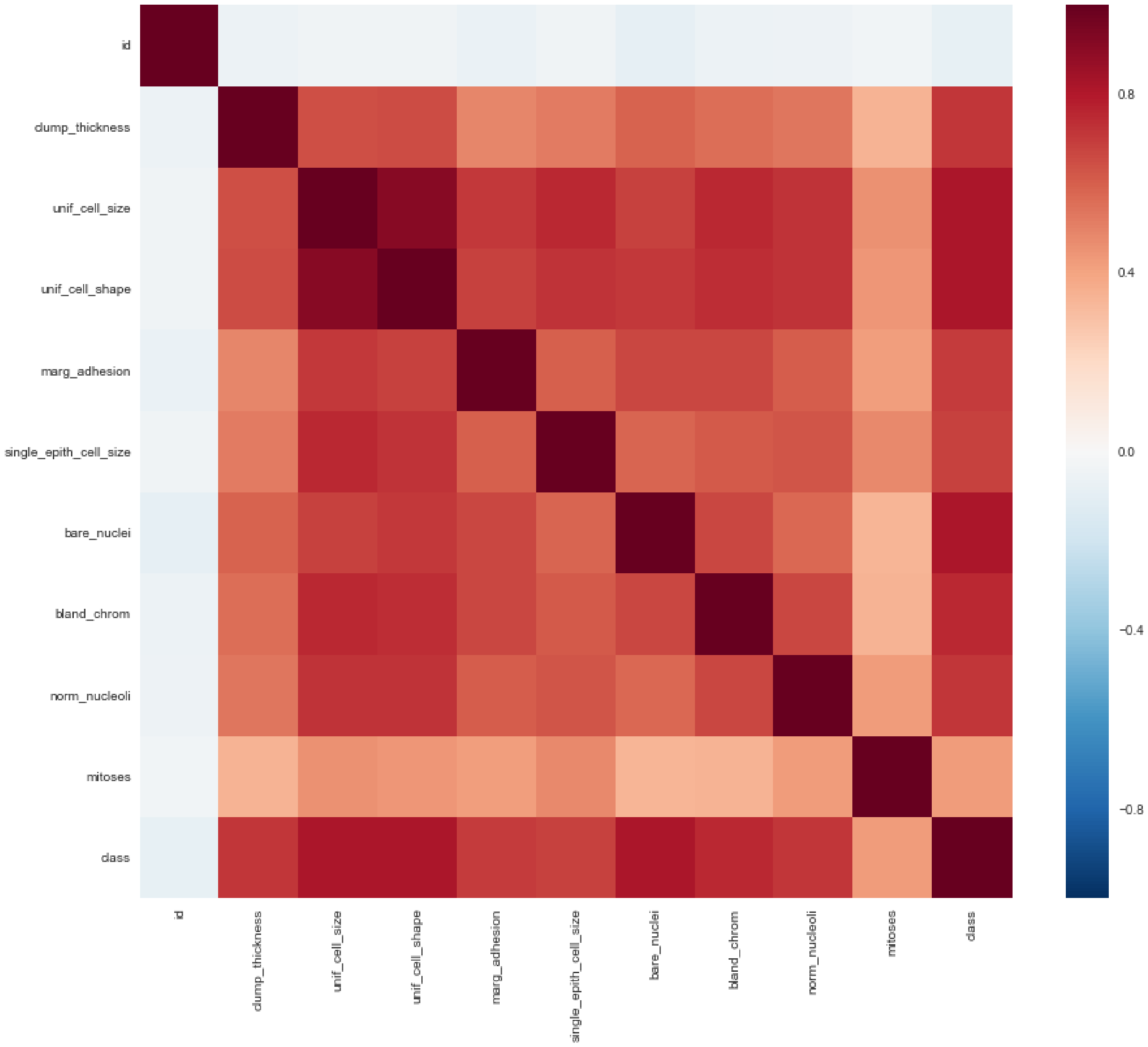


Fig 4. Correlation Analysis

DOI10.17148/IJARCCE.2017.6844

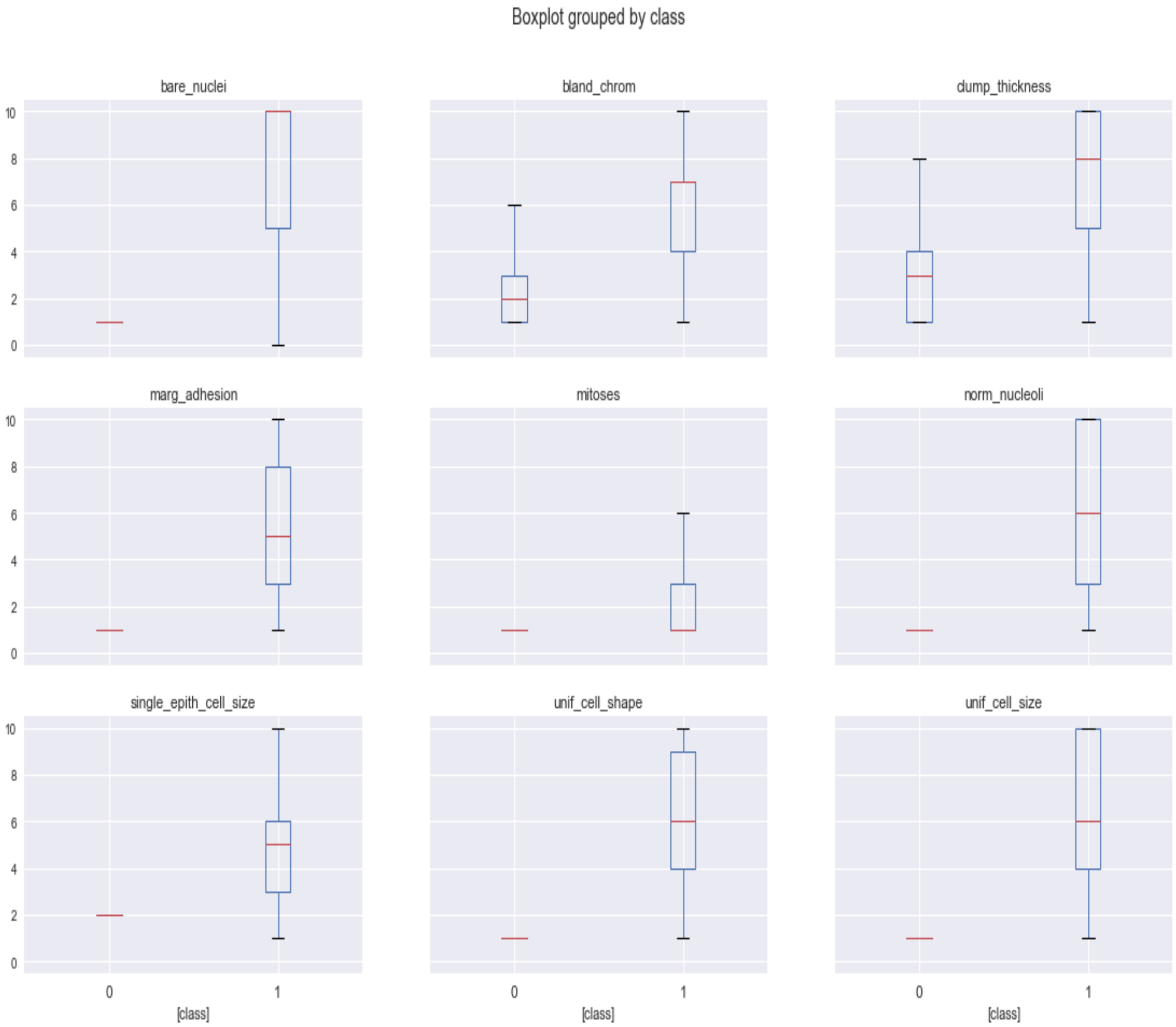
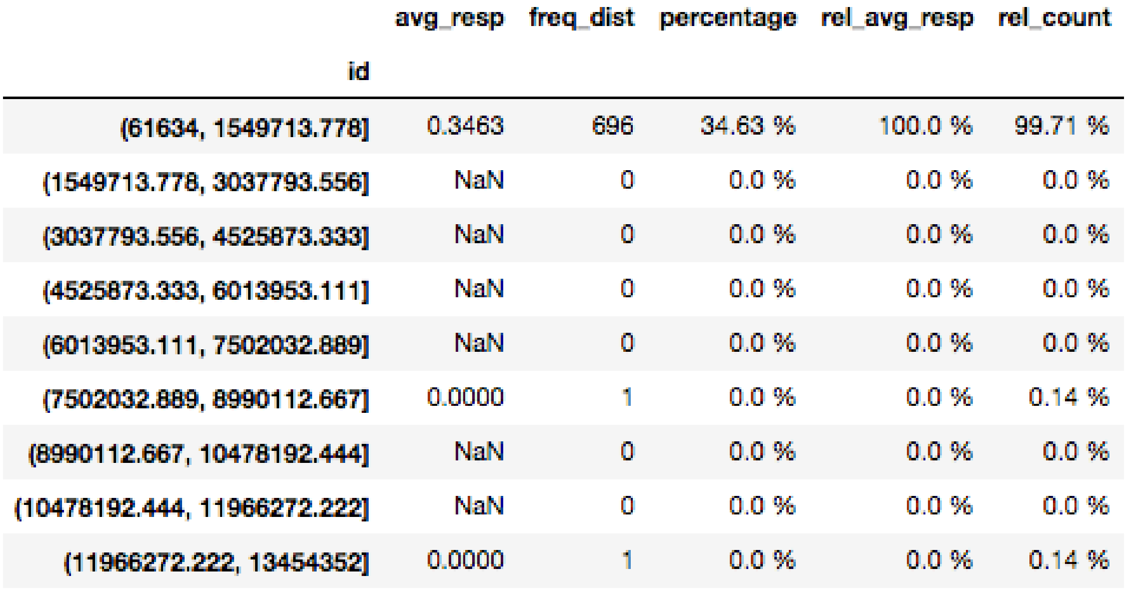


Fig 5. Characteristics grouped by class



id	avg_resp	freq_dist	percentage	rel_avg_resp	rel_count
(61634, 1549713.778]	0.3463	696	34.63 %	100.0 %	99.71 %
(1549713.778, 3037793.556]	NaN	0	0.0 %	0.0 %	0.0 %
(3037793.556, 4525873.333]	NaN	0	0.0 %	0.0 %	0.0 %
(4525873.333, 6013953.111]	NaN	0	0.0 %	0.0 %	0.0 %
(6013953.111, 7502032.889]	NaN	0	0.0 %	0.0 %	0.0 %
(7502032.889, 8990112.667]	0.0000	1	0.0 %	0.0 %	0.14 %
(8990112.667, 10478192.444]	NaN	0	0.0 %	0.0 %	0.0 %
(10478192.444, 11966272.222]	NaN	0	0.0 %	0.0 %	0.0 %
(11966272.222, 13454352]	0.0000	1	0.0 %	0.0 %	0.14 %

Fig6. Relativity Analysis based on id

DOI10.17148/IJARCCE.2017.6844

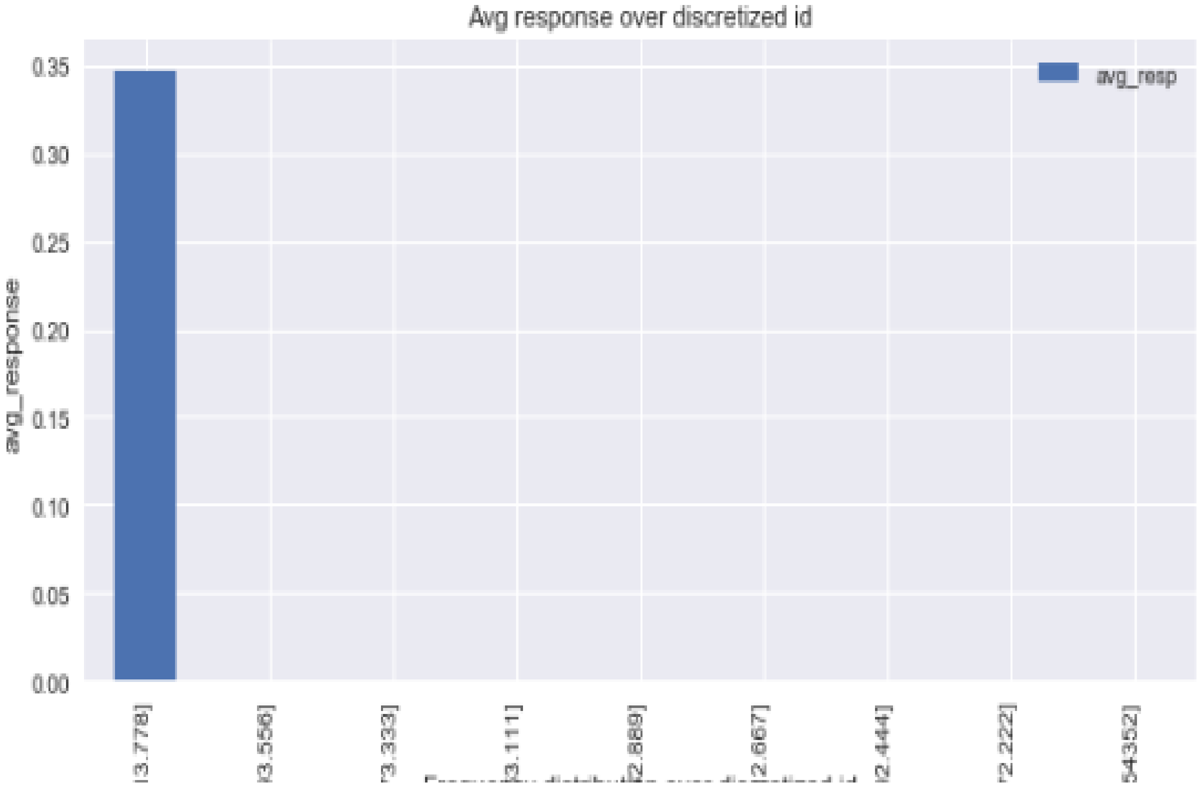


Fig 7. Bar Chart for Average Response related to id

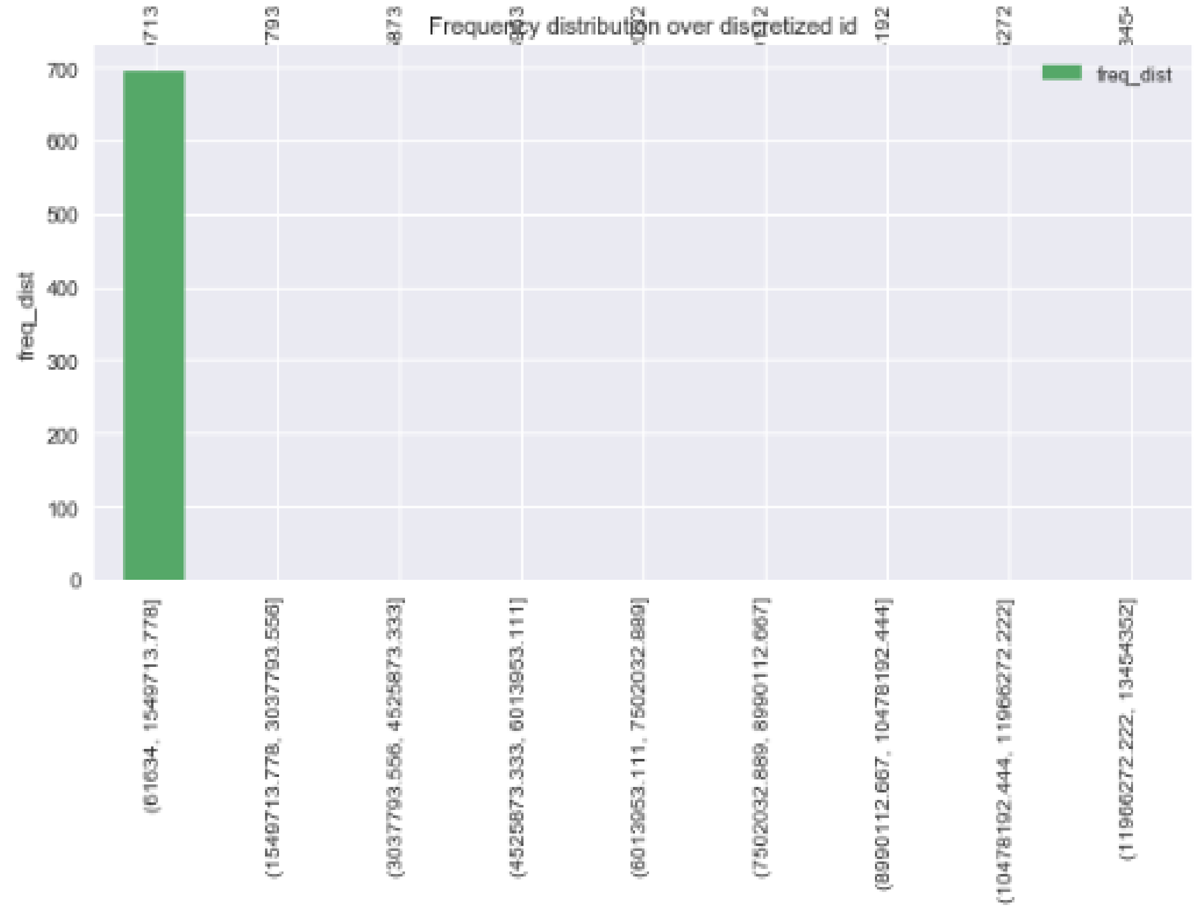
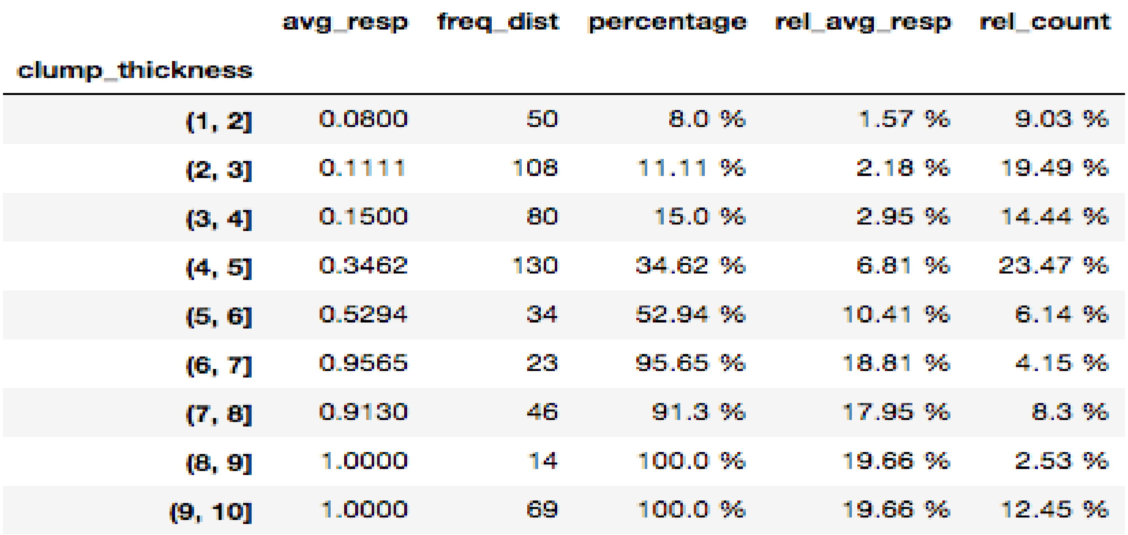


Fig 8. Frequency Distribution over id

DOI10.17148/IJARCCE.2017.6844



clump_thickness	avg_resp	freq_dist	percentage	rel_avg_resp	rel_count
(1, 2]	0.0800	50	8.0 %	1.57 %	9.03 %
(2, 3]	0.1111	108	11.11 %	2.18 %	19.49 %
(3, 4]	0.1500	80	15.0 %	2.95 %	14.44 %
(4, 5]	0.3462	130	34.62 %	6.81 %	23.47 %
(5, 6]	0.5294	34	52.94 %	10.41 %	6.14 %
(6, 7]	0.9565	23	95.65 %	18.81 %	4.15 %
(7, 8]	0.9130	46	91.3 %	17.95 %	8.3 %
(8, 9]	1.0000	14	100.0 %	19.66 %	2.53 %
(9, 10]	1.0000	69	100.0 %	19.66 %	12.45 %

Fig9. Relativity Analysis based on clump_thickness

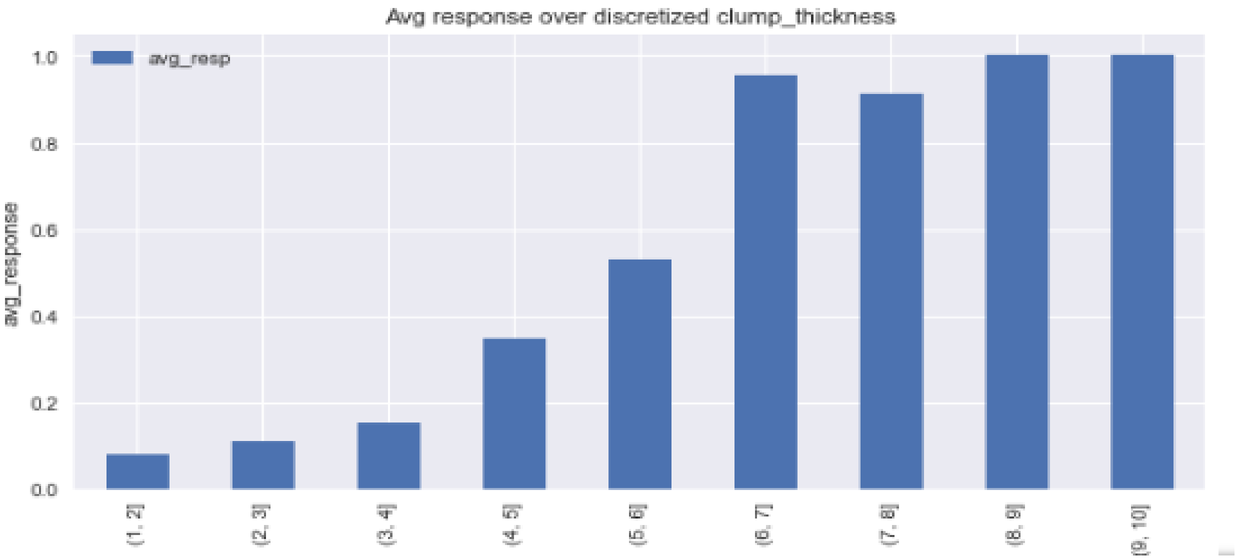


Fig 10. Bar Chart for Average Response related to clump-thickness

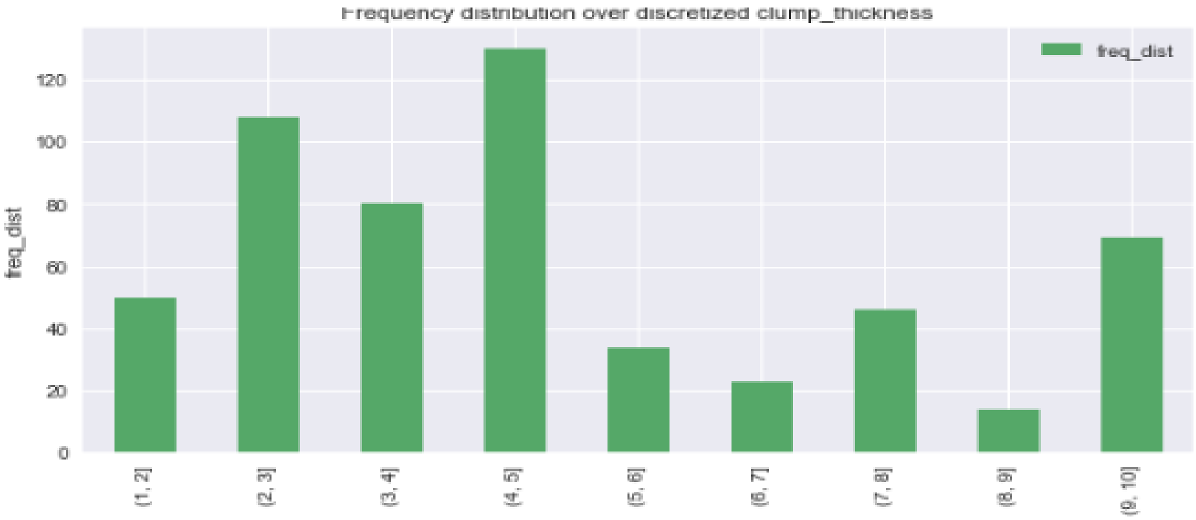
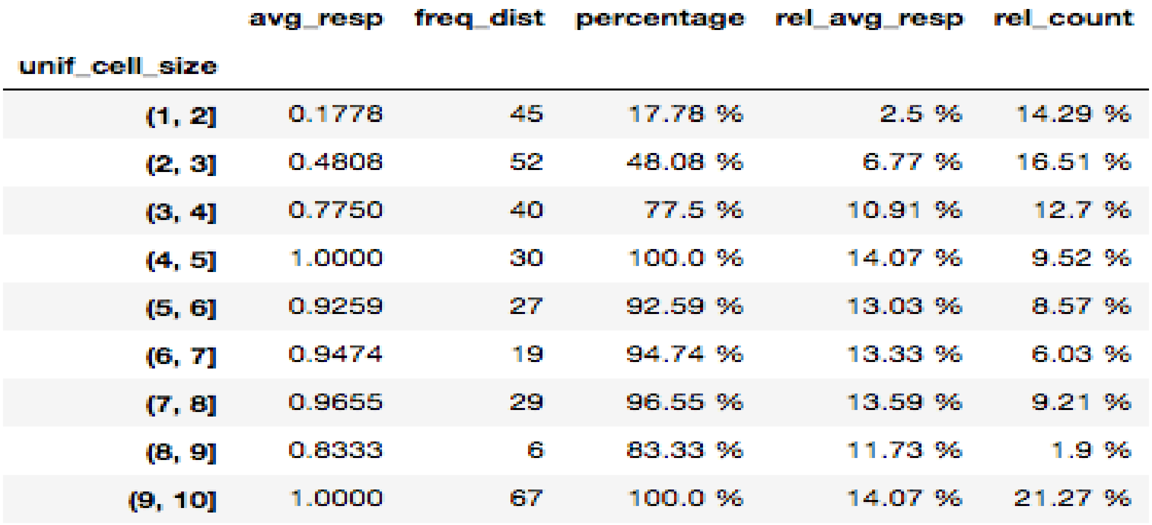


Fig 11. Frequency Distribution over clump-thickness

DOI10.17148/IJARCCE.2017.6844



unif_cell_size	avg_resp	freq_dist	percentage	rel_avg_resp	rel_count
(1, 2]	0.1778	45	17.78 %	2.5 %	14.29 %
(2, 3]	0.4808	52	48.08 %	6.77 %	16.51 %
(3, 4]	0.7750	40	77.5 %	10.91 %	12.7 %
(4, 5]	1.0000	30	100.0 %	14.07 %	9.52 %
(5, 6]	0.9259	27	92.59 %	13.03 %	8.57 %
(6, 7]	0.9474	19	94.74 %	13.33 %	6.03 %
(7, 8]	0.9655	29	96.55 %	13.59 %	9.21 %
(8, 9]	0.8333	6	83.33 %	11.73 %	1.9 %
(9, 10]	1.0000	67	100.0 %	14.07 %	21.27 %

Fig 12. Relativity Analysis based on unif_cell_size

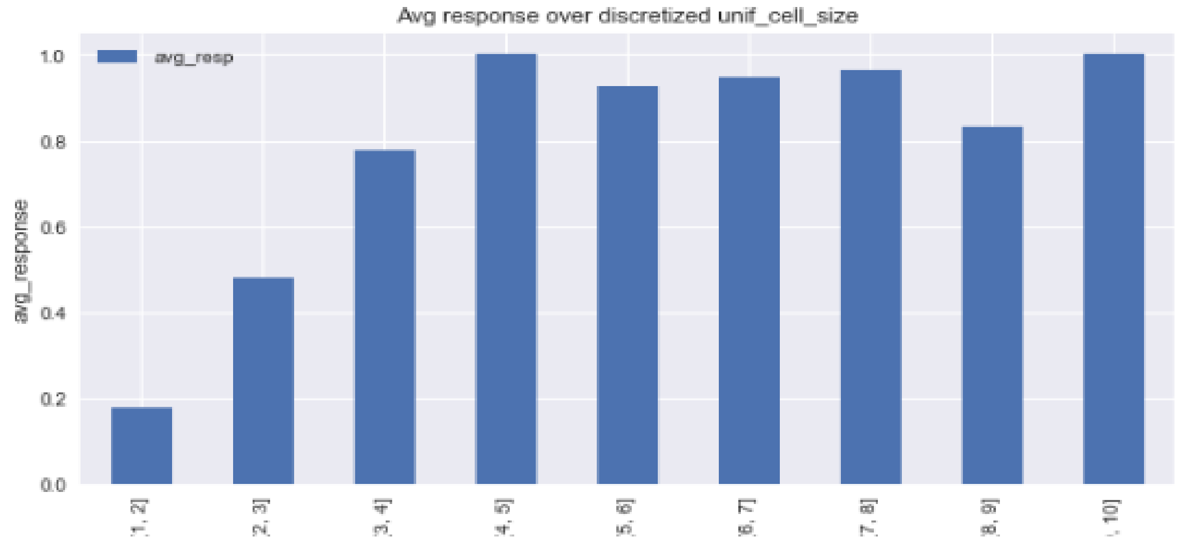


Fig 13. Bar Chart for Average Response related to unif_cell_size

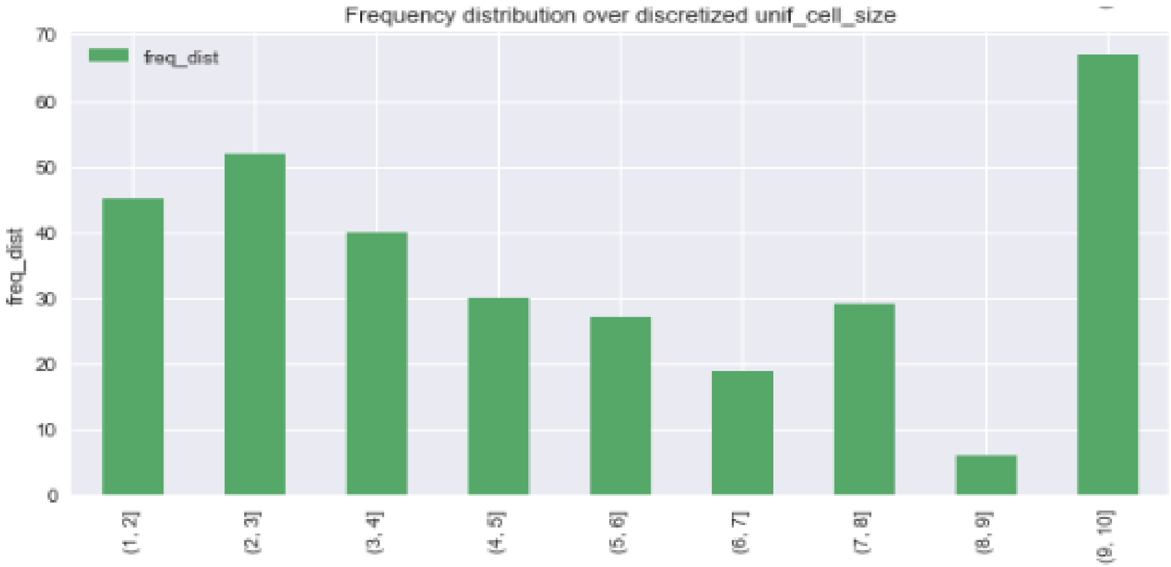
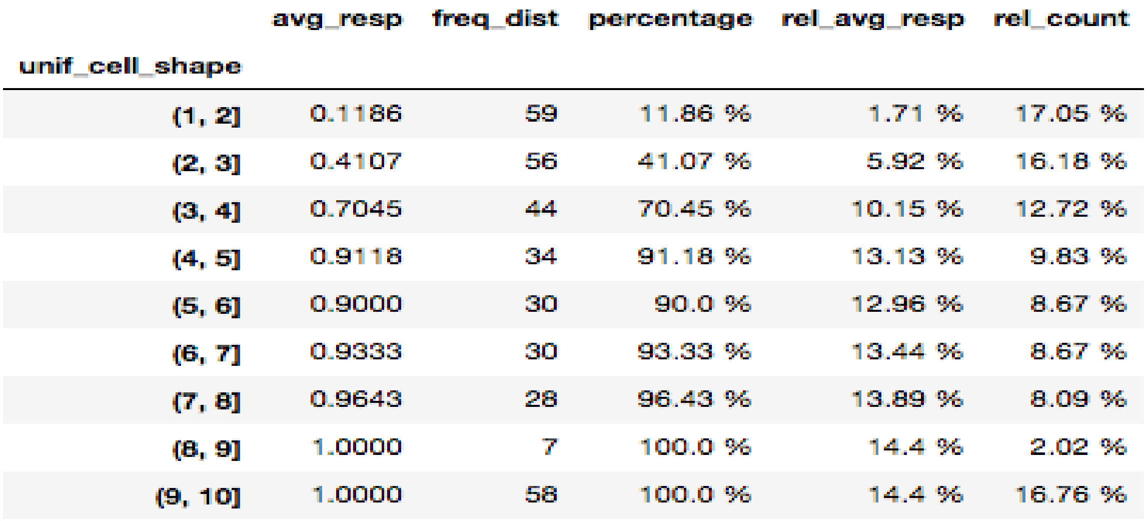


Fig. 14 . Frequency Distribution over unif_cell_size

DOI10.17148/IJARCCE.2017.6844



unif_cell_shape	avg_resp	freq_dist	percentage	rel_avg_resp	rel_count
(1, 2]	0.1186	59	11.86 %	1.71 %	17.05 %
(2, 3]	0.4107	56	41.07 %	5.92 %	16.18 %
(3, 4]	0.7045	44	70.45 %	10.15 %	12.72 %
(4, 5]	0.9118	34	91.18 %	13.13 %	9.83 %
(5, 6]	0.9000	30	90.0 %	12.96 %	8.67 %
(6, 7]	0.9333	30	93.33 %	13.44 %	8.67 %
(7, 8]	0.9643	28	96.43 %	13.89 %	8.09 %
(8, 9]	1.0000	7	100.0 %	14.4 %	2.02 %
(9, 10]	1.0000	58	100.0 %	14.4 %	16.76 %

Fig. 15. Relativity Analysis based on unif_cell_shape

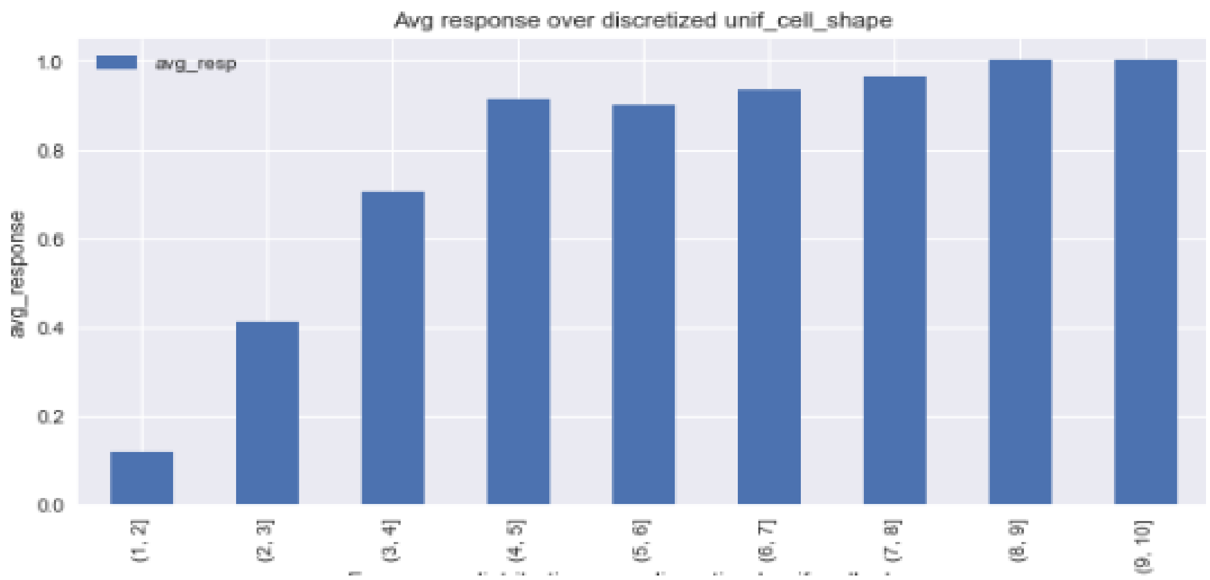


Fig 16. Bar Chart for Average Response related to unif_cell_shape

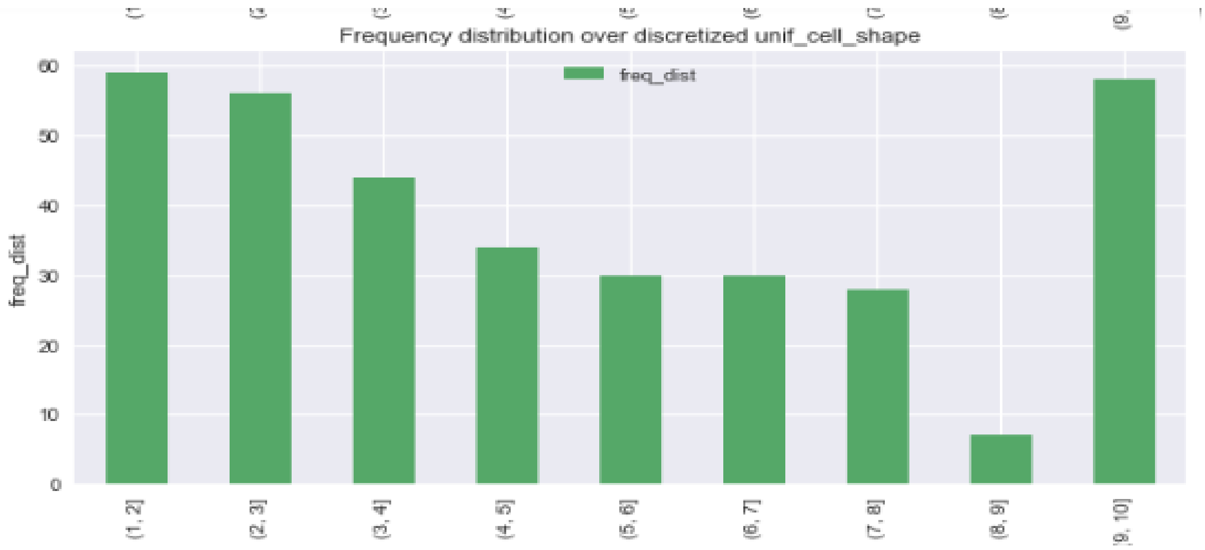
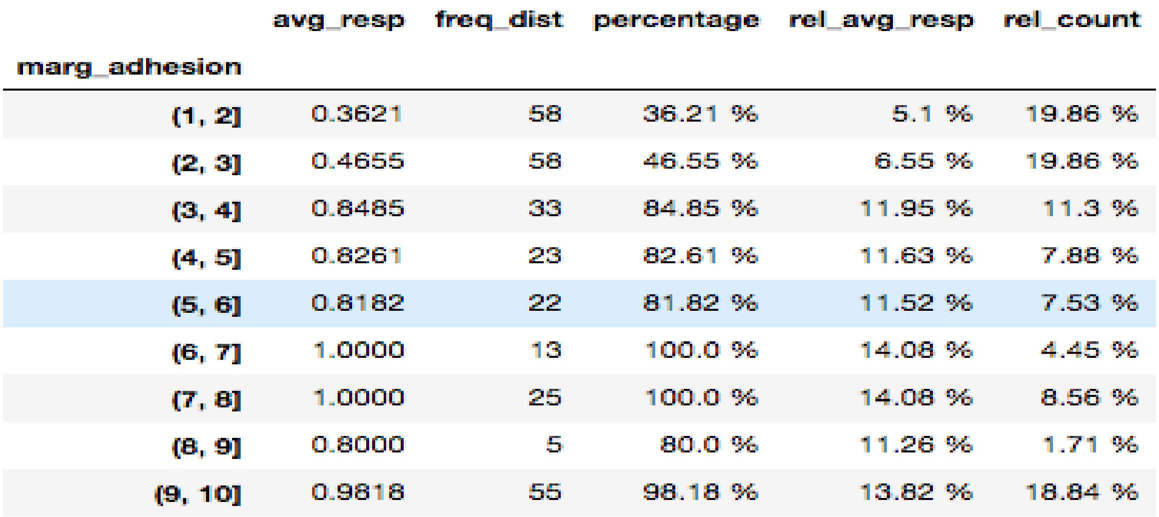


Fig. 17 Frequency Distribution over unif_cell_size

DOI10.17148/IJARCCE.2017.6844



marg_adhesion	avg_resp	freq_dist	percentage	rel_avg_resp	rel_count
(1, 2]	0.3621	58	36.21 %	5.1 %	19.86 %
(2, 3]	0.4655	58	46.55 %	6.55 %	19.86 %
(3, 4]	0.8485	33	84.85 %	11.95 %	11.3 %
(4, 5]	0.8261	23	82.61 %	11.63 %	7.88 %
(5, 6]	0.8182	22	81.82 %	11.52 %	7.53 %
(6, 7]	1.0000	13	100.0 %	14.08 %	4.45 %
(7, 8]	1.0000	25	100.0 %	14.08 %	8.56 %
(8, 9]	0.8000	5	80.0 %	11.26 %	1.71 %
(9, 10]	0.9818	55	98.18 %	13.82 %	18.84 %

Fig 18. Relativity Analysis based on marg_adhesion

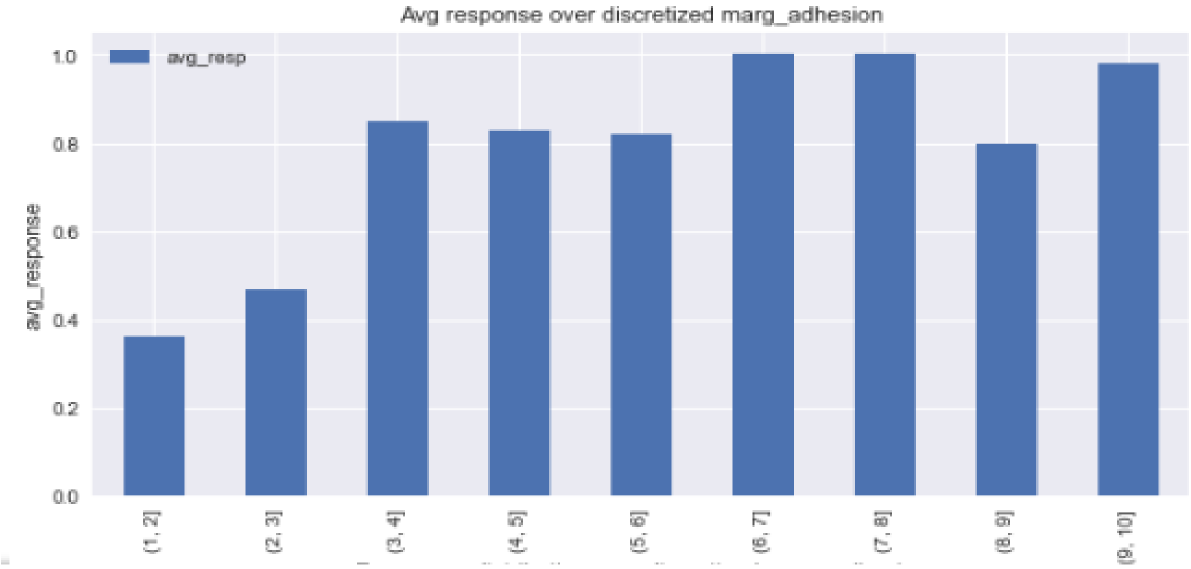


Fig. 19 Bar Chart for Average Response related to marg_adhesion

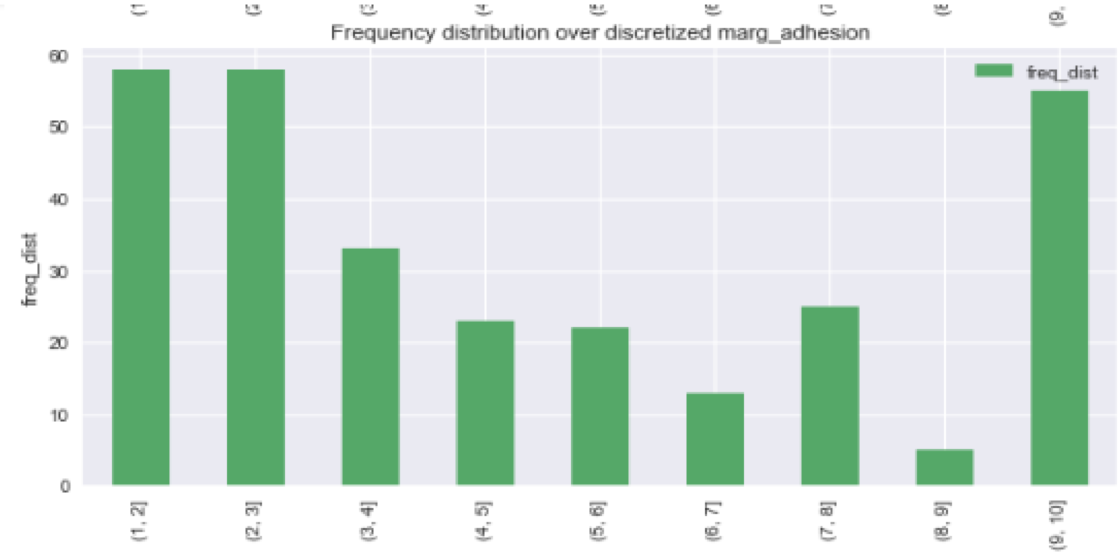
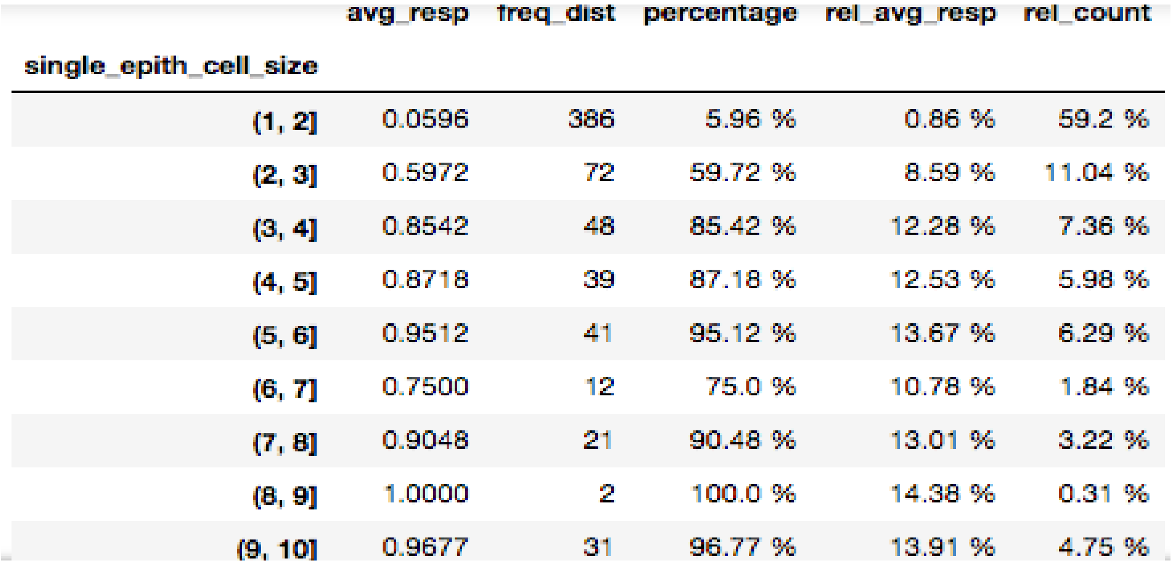


Fig. 20 Frequency Distribution over marg_adhesion

DOI10.17148/IJARCCE.2017.6844



	avg_resp	freq_dist	percentage	rel_avg_resp	rel_count
single_epith_cell_size					
(1, 2]	0.0596	386	5.96 %	0.86 %	59.2 %
(2, 3]	0.5972	72	59.72 %	8.59 %	11.04 %
(3, 4]	0.8542	48	85.42 %	12.28 %	7.36 %
(4, 5]	0.8718	39	87.18 %	12.53 %	5.98 %
(5, 6]	0.9512	41	95.12 %	13.67 %	6.29 %
(6, 7]	0.7500	12	75.0 %	10.78 %	1.84 %
(7, 8]	0.9048	21	90.48 %	13.01 %	3.22 %
(8, 9]	1.0000	2	100.0 %	14.38 %	0.31 %
(9, 10]	0.9677	31	96.77 %	13.91 %	4.75 %

Fig 21. Relativity Analysis based on single_epith_cell_size

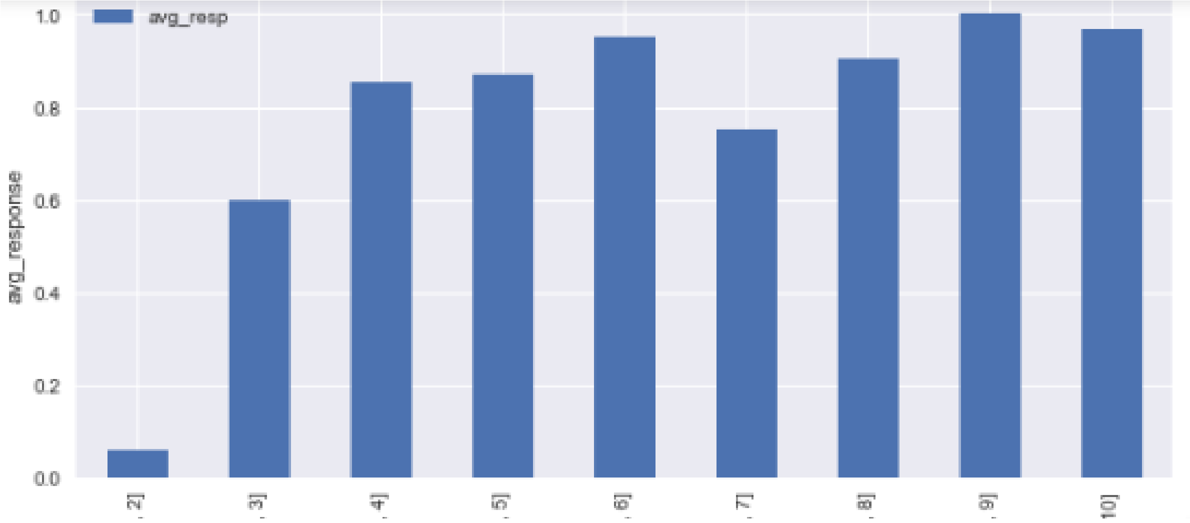


Fig. 22 Bar Chart for Average Response related to single_epith_cell_size

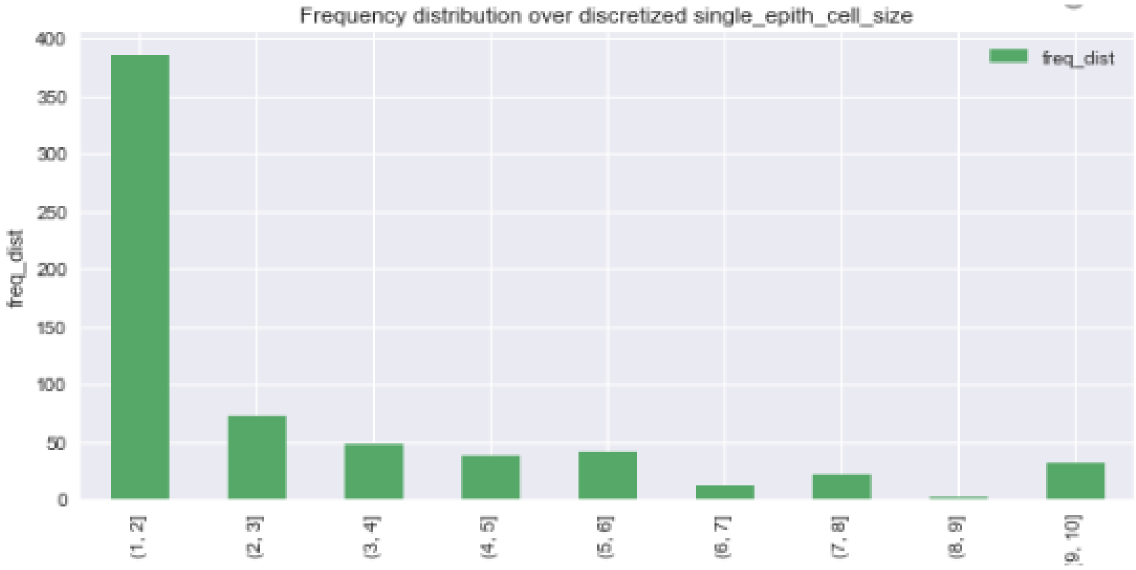
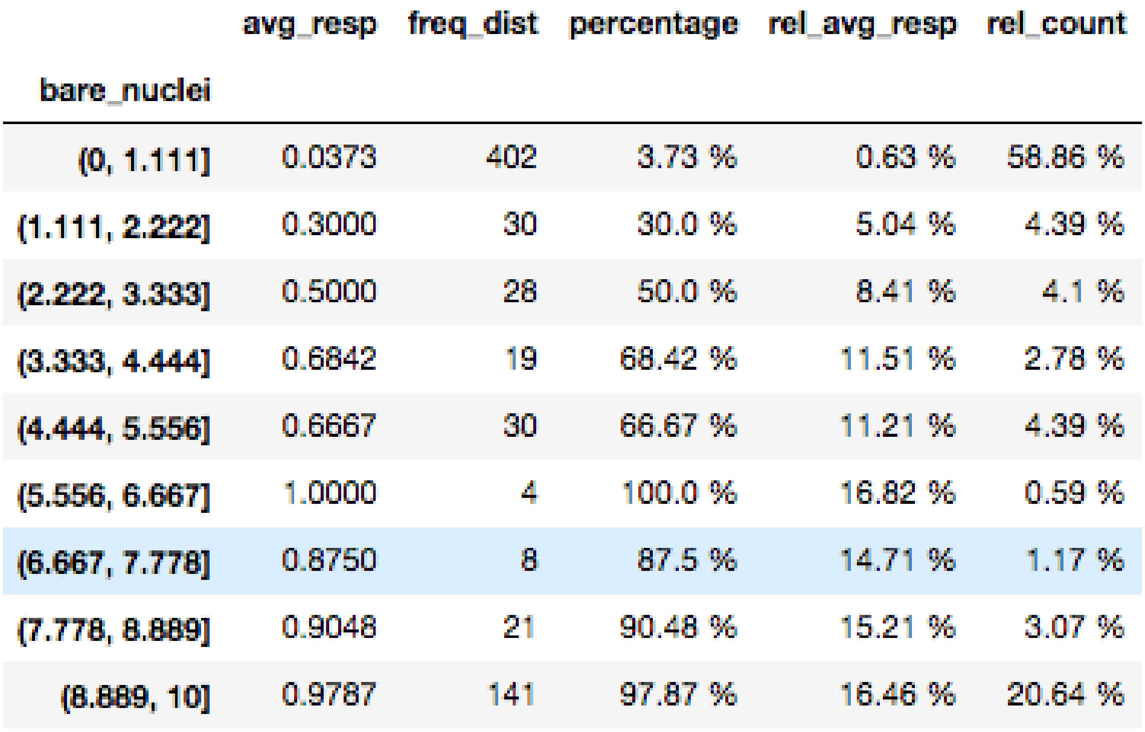


Fig. 23 Frequency Distribution over single_epith_cell_size

DOI10.17148/IJARCCE.2017.6844



bare_nuclei	avg_resp	freq_dist	percentage	rel_avg_resp	rel_count
(0, 1.111]	0.0373	402	3.73 %	0.63 %	58.86 %
(1.111, 2.222]	0.3000	30	30.0 %	5.04 %	4.39 %
(2.222, 3.333]	0.5000	28	50.0 %	8.41 %	4.1 %
(3.333, 4.444]	0.6842	19	68.42 %	11.51 %	2.78 %
(4.444, 5.556]	0.6667	30	66.67 %	11.21 %	4.39 %
(5.556, 6.667]	1.0000	4	100.0 %	16.82 %	0.59 %
(6.667, 7.778]	0.8750	8	87.5 %	14.71 %	1.17 %
(7.778, 8.889]	0.9048	21	90.48 %	15.21 %	3.07 %
(8.889, 10]	0.9787	141	97.87 %	16.46 %	20.64 %

Fig. 24 Relativity Analysis based on bare_nuclei

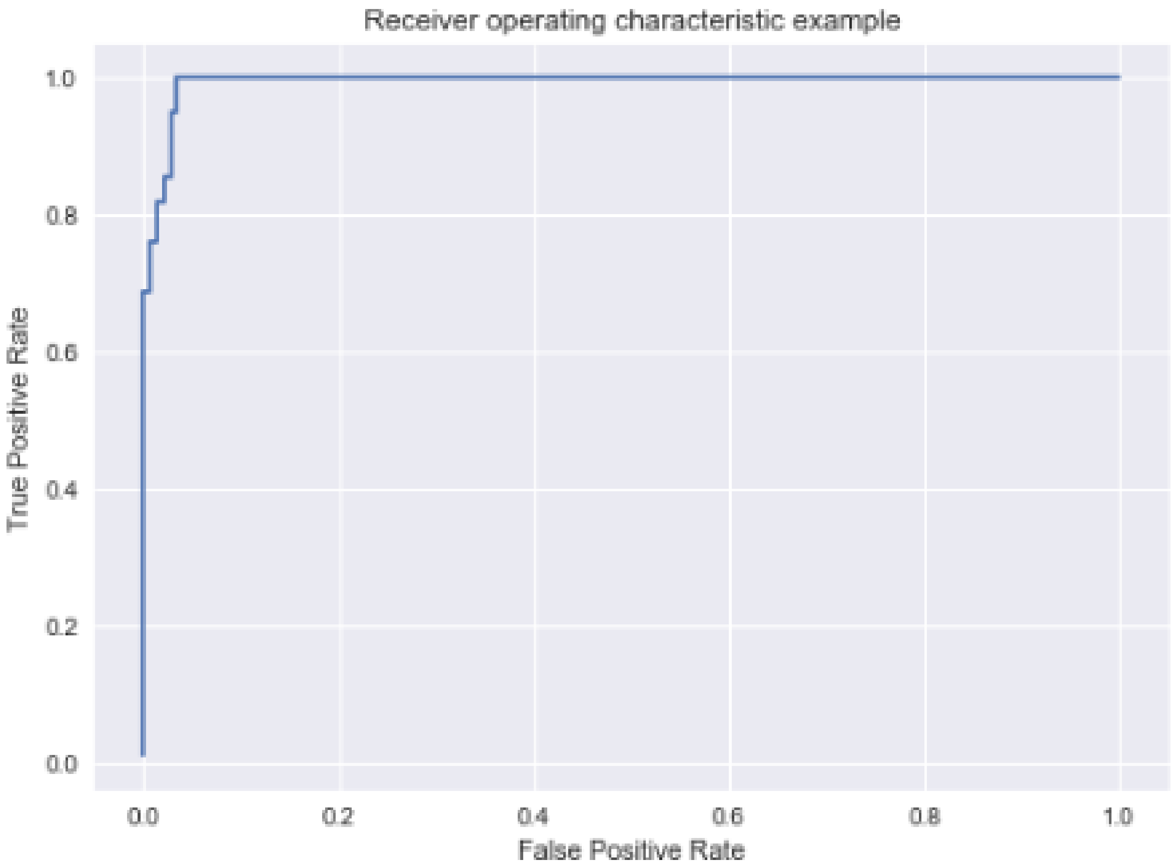


Fig 24. ROC Curve for Logistic Regression

DOI10.17148/IJARCCE.2017.6844

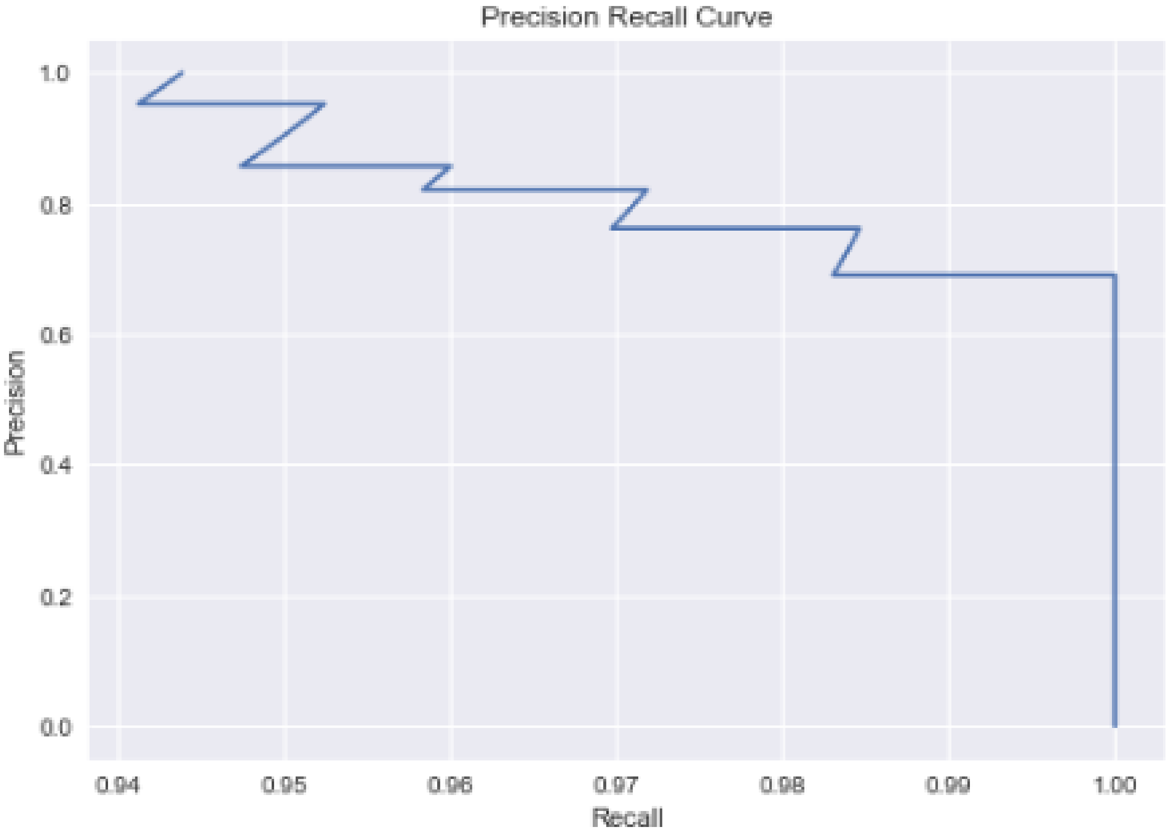


Fig.25 Precision Recall Curve for Logistic Regression

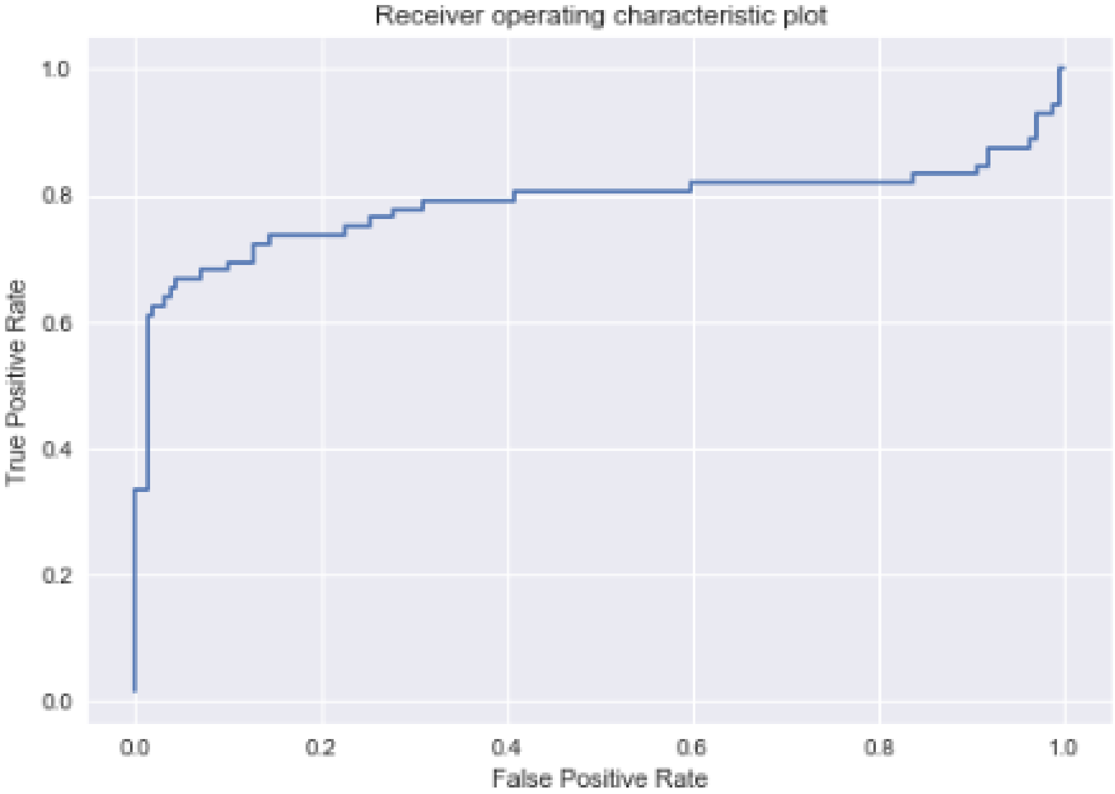


Fig. 26 ROC Curve for Naïve Bayes

DOI10.17148/IJARCCE.2017.6844

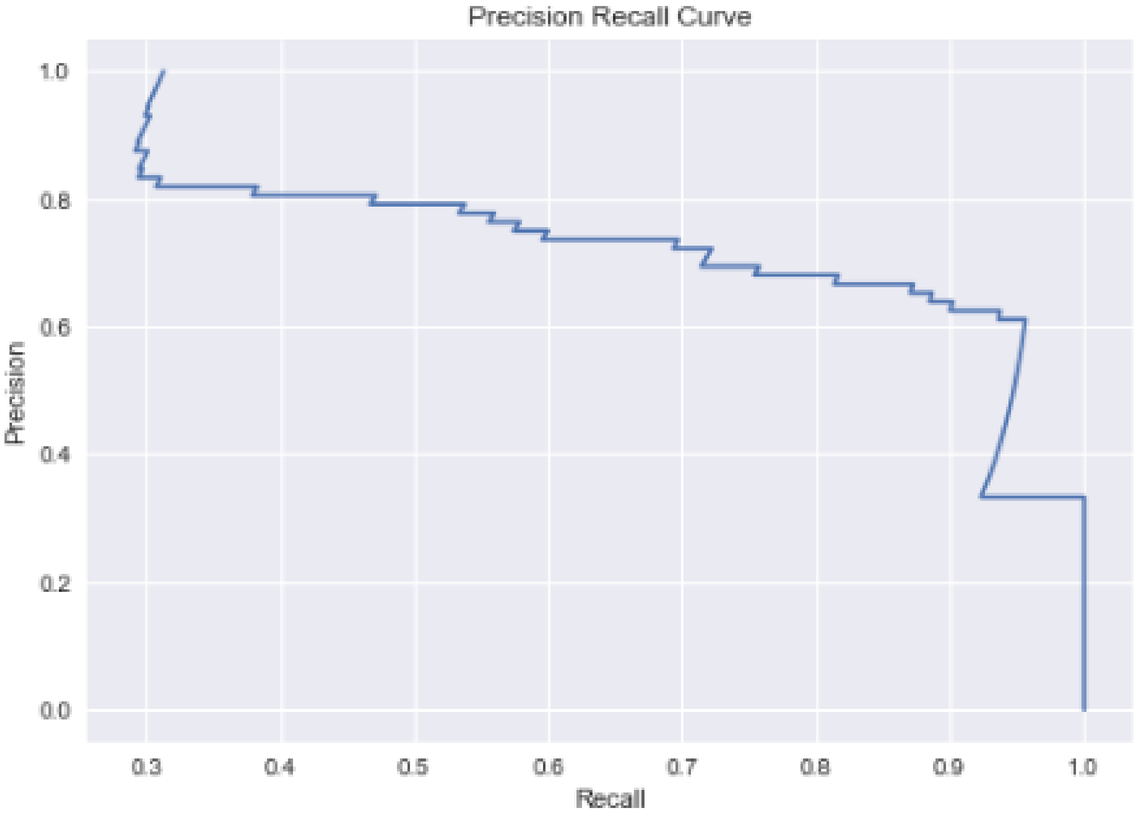


Fig. 27 Precision Recall Curve for Naïve Bayes

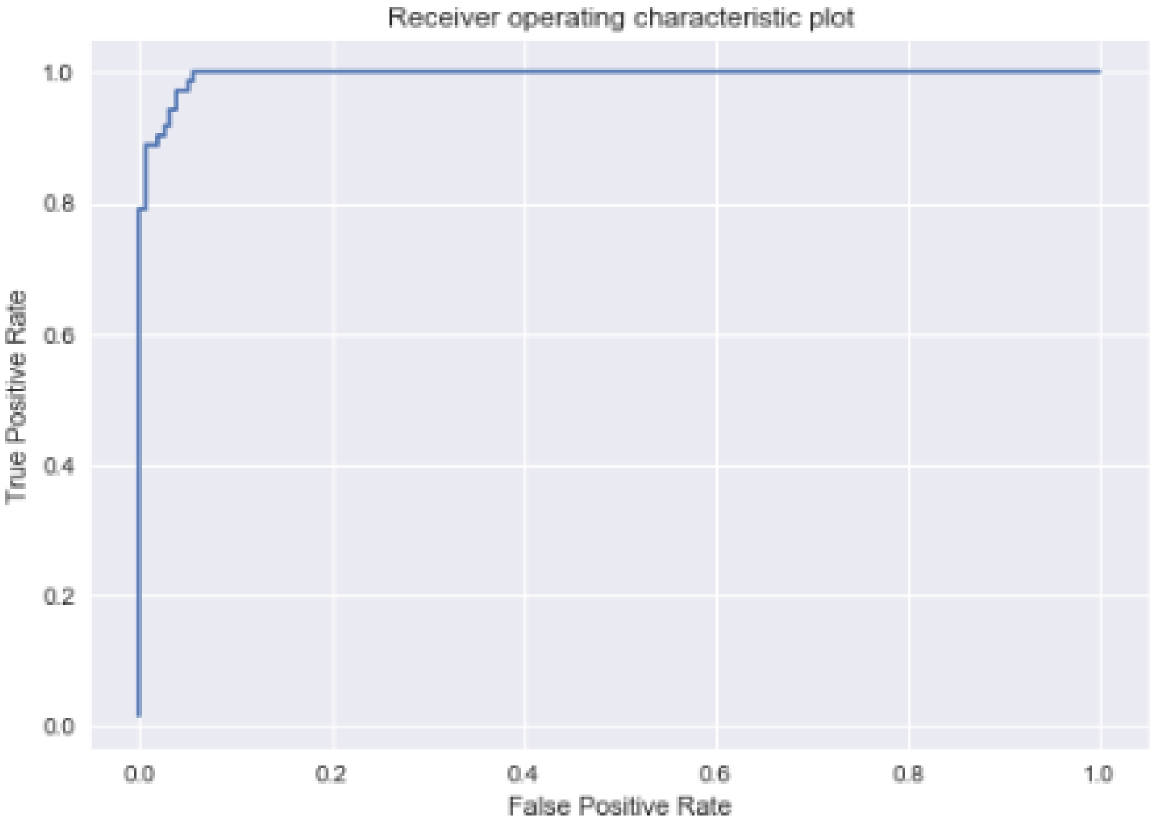


Fig. 28 ROC Curve for Gradient Boosted Trees

DOI10.17148/IJARCCE.2017.6844

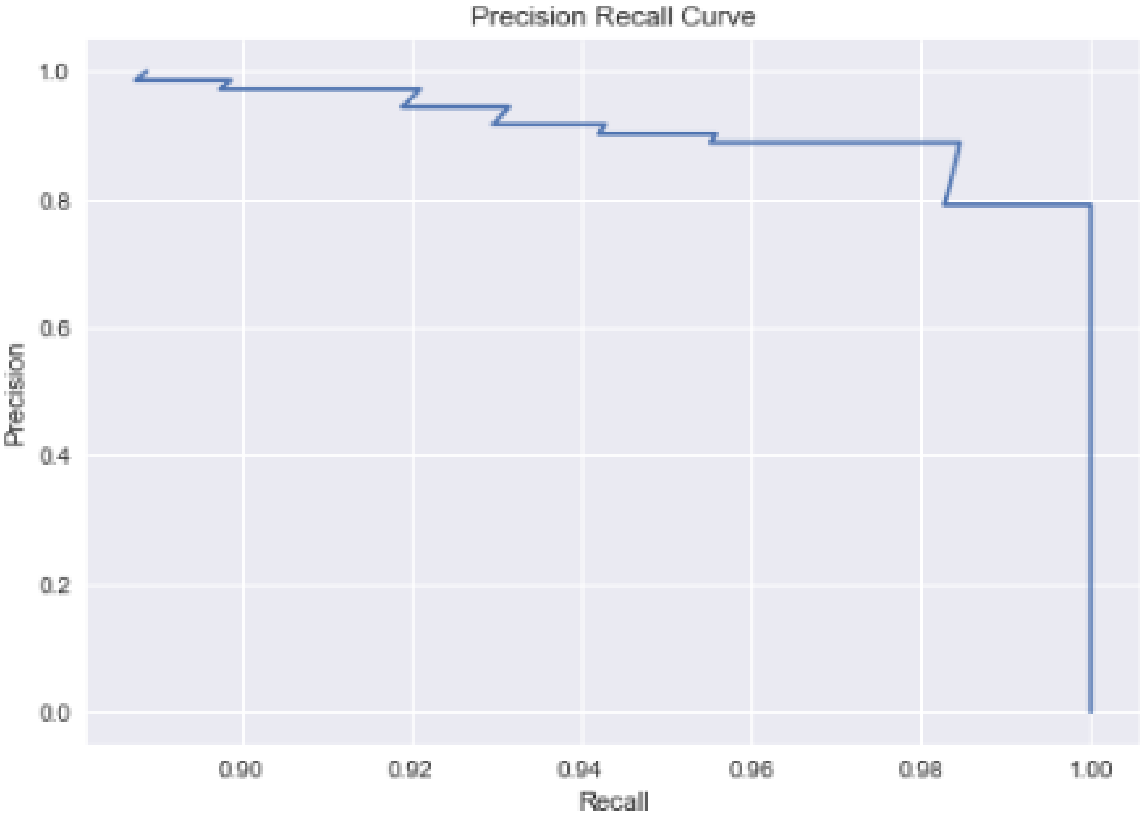


Fig. 29 Precision Recall Curve for Gradient Boosted Trees

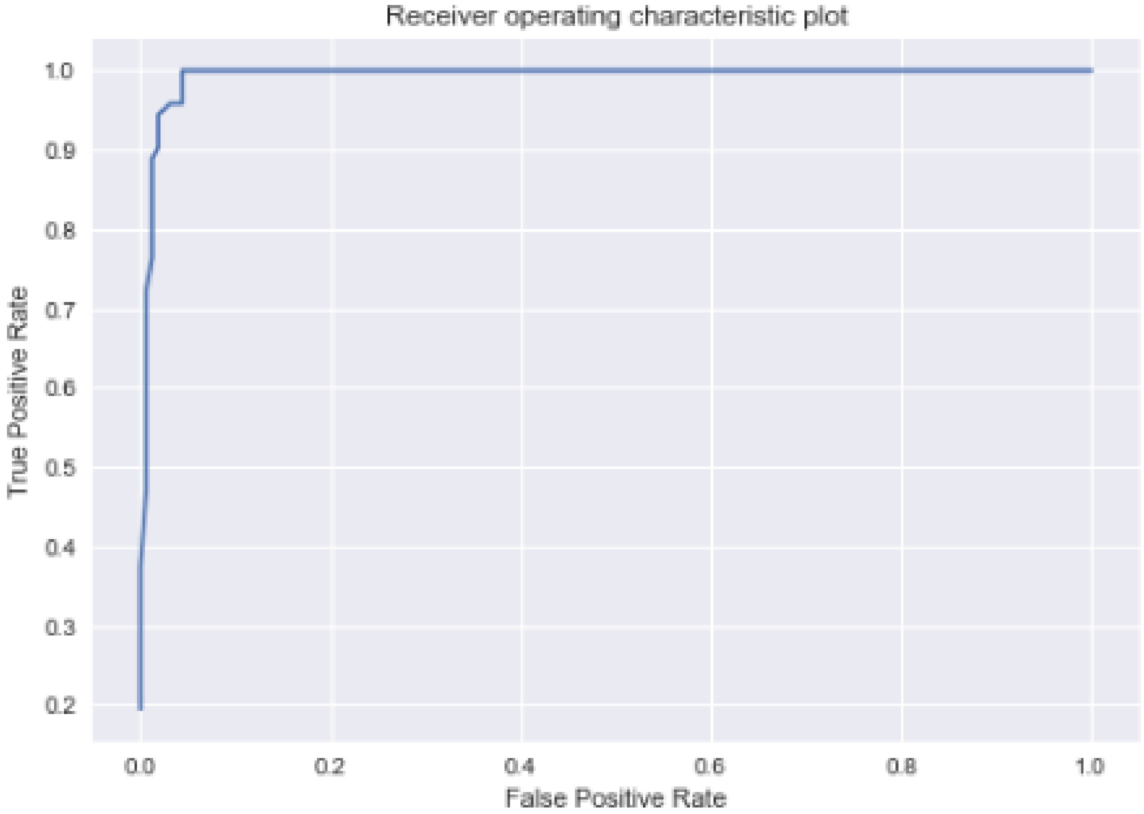


Fig. 30 ROC Curve for Random Forest Trees

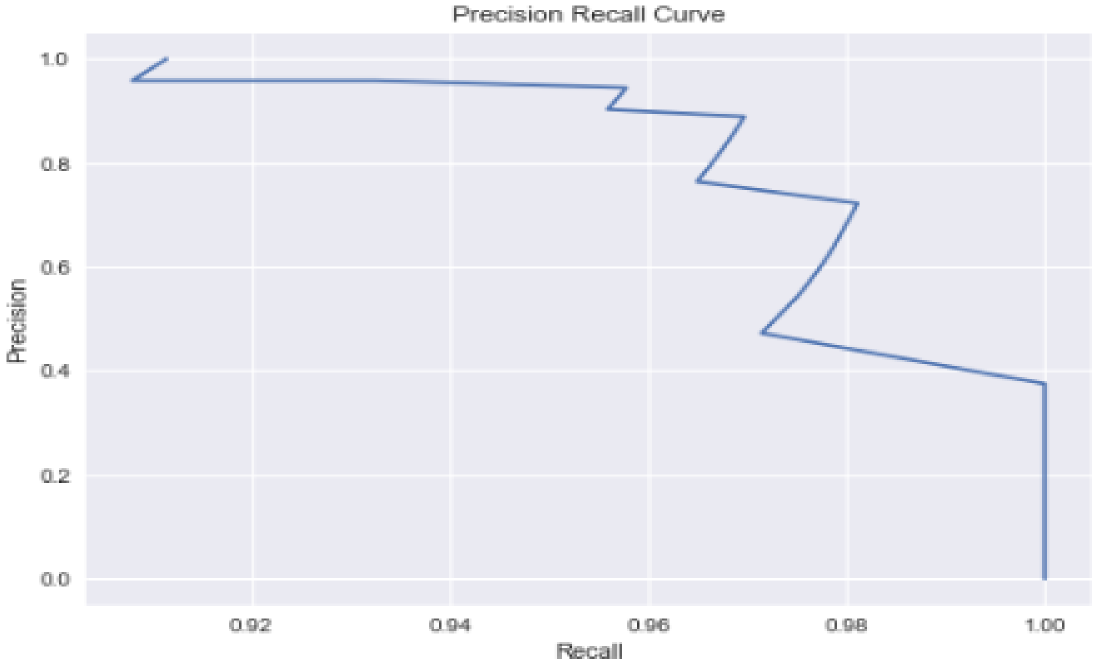


Fig. 31 Precision Recall Curve for Random Forest

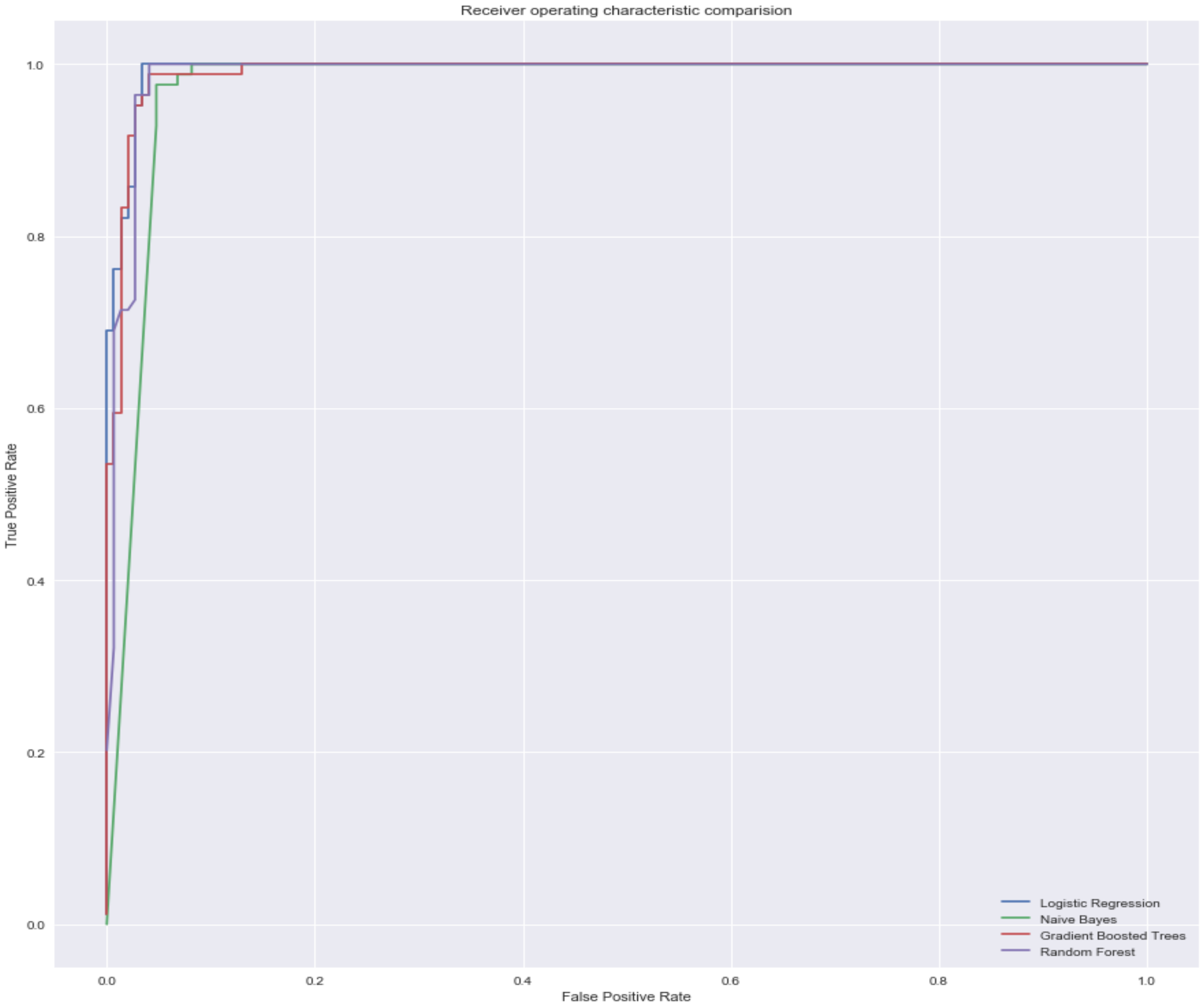


Fig. 32 Comparison Of All Classifiers Using ROC Curve

DOI10.17148/IJARCCE.2017.6844

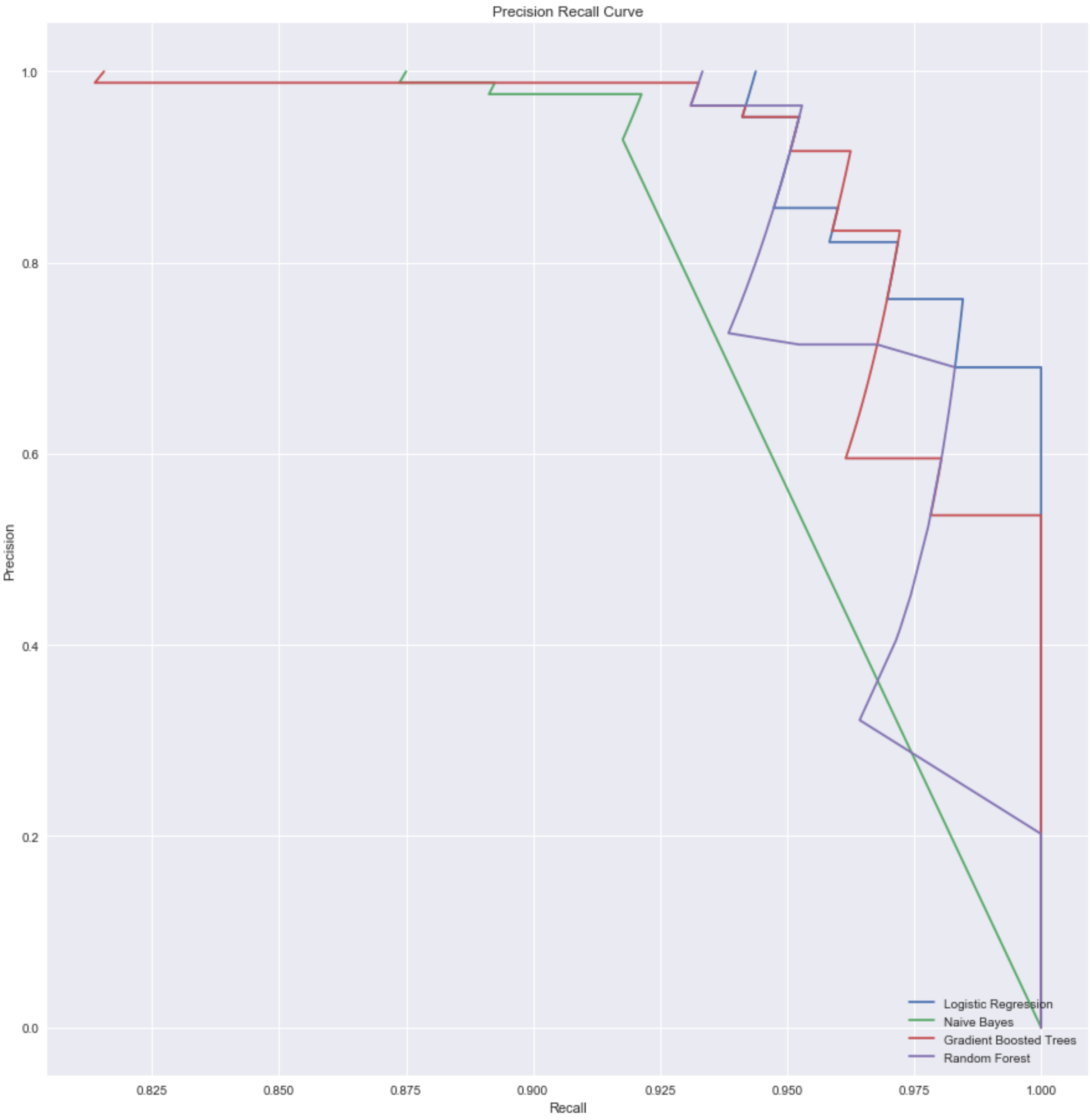


Fig. 33 Comparison of All Classifiers Using Precision Recall

V. CONCLUSION

Since the class balance is not perfect in the dataset, AUC/ROC curve cannot be the sole determiner of the Effectiveness of the model. After evaluating the Precision recall curve along with ROC curve for the algorithms, Logistic regression seems to be a valid model for making predictions, but the choice will be determined by the choice of the operating point; Random forest will be a better choice if the precision requirement is 95%

ACKNOWLEDGEMENT

Every step towards learning requires considerable effort on part of those who wish to make a change. My concentrated effort under the supervision of my Professor **Ms. Shalini L.** Has not only been proved to beneficial for me, it has also motivated me strive for perfection, and for that I offer my sincere gratitude towards her. I would also like to thank my colleagues at work who have offered their insight in this field and have helped me and motivated me in achieving this goal.

DOI10.17148/IJARCCE.2017.6844

REFERENCES

[1] Hosmer, D. W., Jr., & Lemeshow, S. (2000). Applied logistic regression (2nd ed.). New York: Wiley.
[2] Menard, S. (1995). Applied logistic regression analysis (Sage University Paper Series on Quantitative Applications in the Social Sciences, 07–106). Thousand Oaks, CA: Sage.
[3] Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. Biometrika, 78, 691–692.
[4] Tolman, R. M., & Weisz, A. (1995). Coordinated community intervention for domestic violence: The effects of arrest and prosecution on recidivism of woman abuse perpetrators. Crime and Delinquency, 41(4), 481–495.
[5] Lewis, D. David, "Naïve (Bayes) at Fourty," Proceedings of ECML-98, 10th European Conference on Machine Learning, 1998. -
[6] McCallum, Andrew, Kamal Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," In AAAI-98 Workshop on Learning for Text Categorization, 1998.
[7] Applied Predictive Modelling: Kjell Johnson, Max Kuhn
[8] CARAGEA, D., SILVESCU, A.,AND HONAVAR, V. A framework for learning from distributed data using sufficient statistics and its application to learning decision trees. International Journal of Hybrid Intelligent Systems 1, 2 (2004).
[9] QUINLAN, J. R. Induction of decision trees. In Machine Learning (1986), pp. 81–106.
[10] STATISTICS, L. B.,AND BREIMAN, L. Random forests. In Machine Learning (2001), pp. 5–32.

BIOGRAPHIES

Shekhar Tanwar is B.Tech. In computer Science and Engineering from Vellore Institute of Technology and is working as a Network Analyst at Accenture Services. Pvt. LTD. His areas of interest are Big Data, Machine Learning, Data Science, Text Analytics & Natural Language Processing

Prof. Shalini. L is working as Assistant Professor (Senior) at School of computer Science and Engineering, VIT University. Her area of interest are data structure and algorithm, machine learning, natural language processing & Theory or computation. She has been with VIT for more than 10 years and has a total of 18 years of teaching experience.

DOI10.17148/IJARCCE.2017.6844