PROCEEDINGS

2020 IEEE 9th INTERNATIONAL CONFERENCE ON COMUNICATION SYSTEMS AND NETWORK TECHNOLOGIES

Edited by

G.S. Tomar, Chairman, IEEE MP Sub Section, India

Organized by

***IEEE MP Sub Section and Hosted* by**

SRCEM, Banmore, near Gwalior, India

Financial Sponsor: IETE Gwalior, India

Technically Co-Sponsored by

IEEE Bombay Section

Machine Intelligence Research Labs, Gwalior

2020 9th IEEE International Conference on

Communication Systems and Network Technologies

CSNT 2020

Table of Contents

Track C: Mobile Computing, WSN, IOT and next Generation Networks

Track D: Hardware Design, Data Mining and systems

Track E: Cryptography, Blockchain and Security Algorithms

Track F: Image Processing, Signal Processing intelligent systems

Demystifying and Anticipating Graduate School Admissions using Machine Learning Algorithms

Mohd Aijaj Khan , Manish Dixit
Madhav Institute of Technology and Science
Gwalior, India
AijajKhan@ieee.org, Manishdixit@ieee.org

Aaradhya Dixit
Electronics & Communication Engg, Birla Institute of Technology, Mesra, India,
AaradhyaDixit@ieee.org

***Abstract*—One of the many aspirations of undergraduate students in India is going for further graduate studies. Unfortunately, many students spend months and years of preparation focusing on things that unfortunately won't improve their chances of getting into a good graduate school. This paper evaluates the chances of applicants to get into a particular graduate program using various classification and regression approaches of Machine Learning. Various algorithms have been pitted against each other and also the most important features have been extracted which are useful to get into a graduate school program. Using unsupervised approach, this paper finds various categories of students and pool them together to find if they are perfect fit for admission or not. A novel approach of predicting the chances for admission in graduate school is introduced in this paper.**

Keywords—Machine Learning, Support Vector Machines, Artificial Neural Networks, Elbow Method, Keras, Python, Scikit, Pandas, Features, Redundancy, Supervised Learning, Unsupervised Learning, Cumulative Grade Point Average (CGPA), Graduate Record Examinations (GRE), Test of English as a Foreign Language (TOEFL), Accuracy, F Score, R^2 Score, Hidden Layers, Confusion Matrix, Epoch, Metric, Regression, Classification.

I. INTRODUCTION

The complicated procedure of applying to a graduate school in the USA results in a very hectic undergraduate schedule for Indian students. The parallel approach of keeping a high Cumulative Grade Point Average, with good GRE, TOEFL scores and publishing research papers, getting good Letters of Recommendation and making a good Statement of Purpose certainly makes every Indian student busy who wants to further excel in academia. [1] We understand that all the things cumulatively are very important for graduate school admissions. But what are the most important factors? If we can find those most important factors, we can emphasize more on them to increase our chances to get an admit. Given a labeled dataset of 500 students who applied to a graduate program, we will find the machine learning algorithm which will very closely predict the chances of admission. [4] And from these techniques, we will also extract some of the redundant and very important features. This paper also takes an approach to find the relation of the features for evaluating the chances of admit from a graduate school. We will also convert it into a classification problem and similarly evaluate a confusion matrix with the aid of classification algorithms and the dataset. Several Machine Learning techniques have been used here and comparative analysis on results of every approach has been done to formulate a novel approach to predict the probability of admission. Also, many powerful techniques such as Support Vector Machines and Artificial Neural Networks have also been used to predict the same. [3]

II. APPROACHES USED

Given a dataset of 500 students, who had applied to the same graduate program earlier, we see the trend of the features, try to manipulate them, remove redundant features and fit them into models for predicting the chances of admission. [4] Labels are also converted into binary format for classification. The models used are as follows. [2][3]

A. *Regression Algorithms*

- Linear Regression
- Support Vector Regression [7]
- Ridge Regression
- Bayesian Ridge Regression
- Artificial Neural Networks
- Random Forest Regression [6]
- Ada Boost Regression
- K-Nearest Neighbors Regression
- Decision Tree Regression

B. *Classification Algorithms*

- Artificial Neural Networks
- Logistic Regression
- Naïve Bayes Classifier
- Support Vector Machines [7]
- Perceptron
- K-Nearest Neighbors Classifier
- Random Forest Classifier [6]
- Decision Tree Classifier

We will also cluster our data into various clusters by K-Means Clustering for distinguishing the low-performing applicants from the outstanding ones to help them to prepare more to achieve their goals. We will be using Python as our programming language along with Numpy, Scikit-learn, Keras for Machine Learning, Pandas for data

978-1-7281-4976-9/20/$31.00 ©2020 IEEE
DOI: 10.1109/CSNT.2020.05

manipulation and Matplotlib and Seaborn libraries for data visualization and analyzing our learning algorithms. [5]

III. MACHINE LEARNING AND FEATURE MANIPULATION

Machine Learning is a way to train our machines to learn certain things on their own without programming them explicitly for every task. The machine learns using the learning algorithms, do some task repeatedly and learn from their experience. For Machine Learning problems, first we must analyze our features, plot the input variables against the target and see which features are important and which ones are redundant. We will also see that if there is also any feature that is looking unimportant but is essential for getting our data fit. We will also try to map correlation between different factors that will help to analyze our dataset. [8]. First, we will import our dataset by using Pandas, and we will see the columns and for any missing values. Columns are as follows with the range of their values:

- Serial No. (1-500)
- GRE Score (260-340)
- TOEFL Score (0-120)
- University Rating (1-5)
- SOP (1-5)
- LOR (1-5)
- CGPA (0-10)
- Research (0 or 1)
- Chance of Admit (0 to 1)

Here, "Chance of Admit" is our label (target value) whose value is between 0 and 1 (Probability of getting admitted). Serial No. is a redundant field and we don't require it, so we will be dropping it from our dataset. CGPA has a mean of 8.57 and its Standard deviation is 0.6. GRE Score has a mean of 316 and its Standard deviation is 11. TOEFL Score has a mean of 107 and its Standard deviation is 6. Which means the scores of the applicants do not differ by much. Also, our Research (value 1 or 0) has a mean of 0.56 which means 56% of applicants have done research work.

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.00000	500.000000	500.000000	500.00000
mean	250.500000	316.472000	107.192000	3.114000	3.374000	3.48400	8.576440	0.560000	0.72174
std	144.481833	11.295148	6.081868	1.143512	0.991004	0.92545	0.604813	0.496884	0.14114
min	1.000000	290.000000	92.000000	1.000000	1.000000	1.00000	6.800000	0.000000	0.34000
25%	125.750000	308.000000	103.000000	2.000000	2.500000	3.00000	8.127500	0.000000	0.63000
50%	250.500000	317.000000	107.000000	3.000000	3.500000	3.50000	8.560000	1.000000	0.72000
75%	375.250000	325.000000	112.000000	4.000000	4.000000	4.00000	9.040000	1.000000	0.82000
max	500.000000	340.000000	120.000000	5.000000	5.000000	5.00000	9.920000	1.000000	0.97000

Fig 1. Information of our dataset including median, standard deviation and mean

What we infer from the data is that, GRE scores are directly proportional to the chance of admit. And so, are TOEFL and CGPA. The Standard Deviation for GRE is 11.29 and TOEFL is 6.08. The Scores of applicants do not differ by much. We will boxplot GRE, TOEFL and CGPA scores to see the distribution, mean, median of scores and outliers, if any

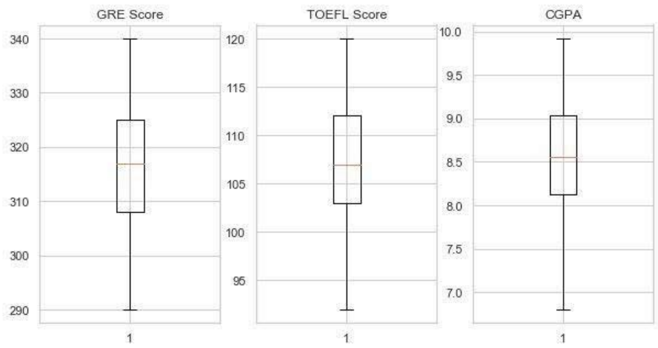


Fig 2. Box Plot to see any abnormalities in the distribution of features.

To find the correlation between different features, we will plot the heatmap of correlation matrix where darker shades denote a large correlation and lighter ones means they are less dependent on each other. The most important column here is to see the "Chance of Admit" to find which factors weigh more for admission.

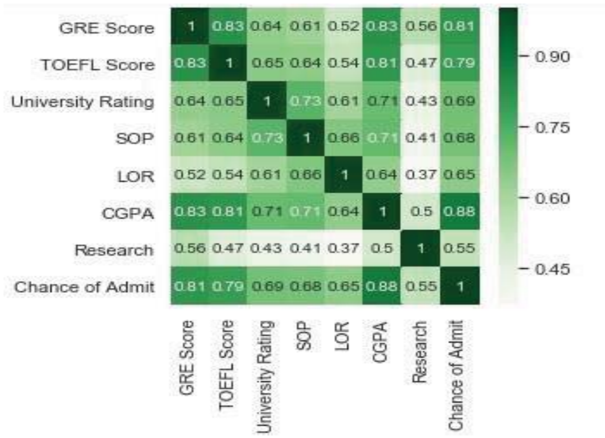


Fig 3. Correlation of different features of dataset using heatmap

We find here that chance of admit has a high correlation with CGPA, GRE and TOEFL scores. Taking the values, we get as expected:

GRE Score	0.810351
TOEFL Score	0.792228
University Rating	0.690132
SOP	0.684137
LOR	0.645365
CGPA	0.882413
Research	0.545871
Chance of Admit	1.000000

Now, we will plot the curve between CGPA and

Chance of Admit and we will find that the relation between them is linear

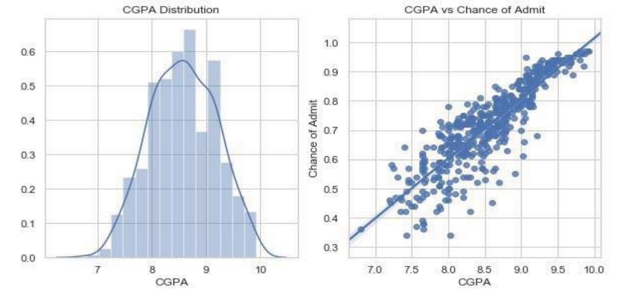


Fig 4. Distribution of CGPA and the Linear Relation of CGPA vs Chance of Admit

Also, the relation between GRE Scores and target variable is linear and the same holds true for CGPA. Hence, the assumptions are verified. The distribution graphs for GRE scores are also plotted. The same is applicable for TOEFL scores according to dataset. [5]

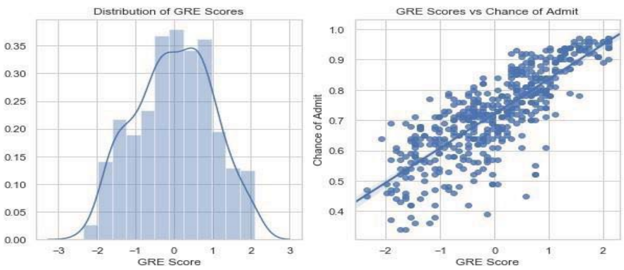


Fig 5. Distribution of GRE scores and the plot of GRE against Chance of Admit (Linear)

We know that Research Experience also plays a very important role in Grad School admissions. Recall from earlier that only 56% of the applicants had research experience.

Total Students: 500

Students having research: 280

Students not having research: 220

Percentage of students having research: 56.0%

The factor plot of Research against Chance of Admit and the distribution graphs are shown in Figure 6.

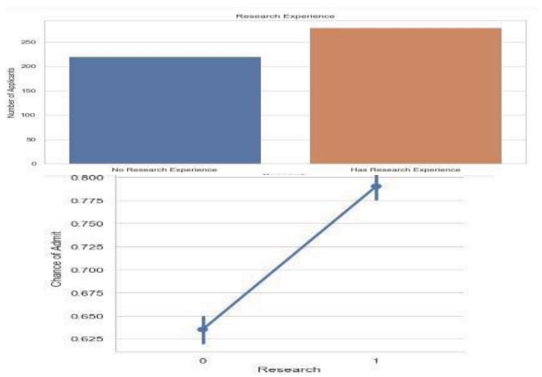


Fig 6. Count of Applicants by Research done and Factor Plot of Research on Admit

And, University rating also plays a very big role. Say, an applicant from IIT having same credentials as compared to an applicant from a Tier 3 university will have a higher chance of getting admission into the Graduate school. For this dataset, the higher the rating, the lower the Tier. Example- University Rating of 5 means Tier 1 university, which is the best of all tiers of universities. The number of students who applied from various tiers of universities are plotted in the next graph shown in Figure 7.

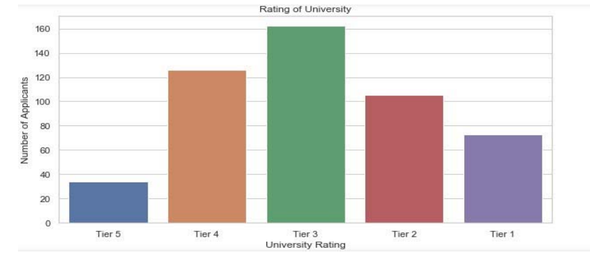


Fig 7. Number of Applicants from different tiers of universities.

The chance of admit vs rating is also given below in Fig 8:

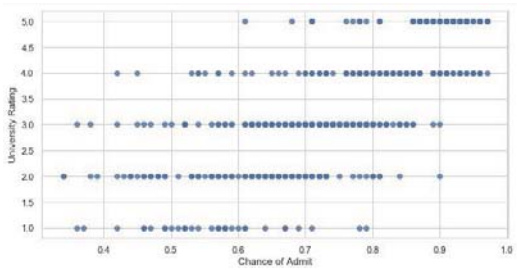


Fig 8. Chance of Admit given University Rating

Letters of Recommendation are being seen redundant here. But, they are related to interaction between professors and students under whose supervision they have performed research work. [5] The students who have done research must have more chances of admission. They are plotted in Fig.9:

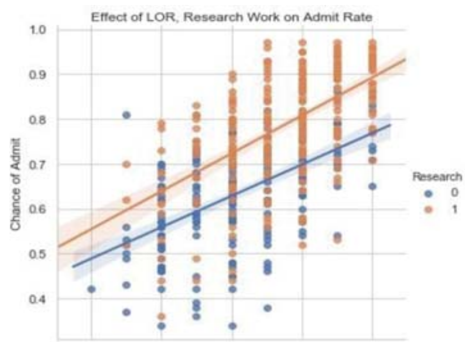


Fig.9: Effect of LOR in conjunction with research work on Chance of Admission

Now we will perform mean normalization of the columns of GRE and TOEFL as they have high ranges. [8] The formula is:

$$X_{i}\_new = (X_i - Mean) / Standard\ Deviation \quad \ldots(1)$$

IV. APPLYING MACHINE LEARNING REGRESSORS AND OBSERVING RESPONSE

Now we will apply various machine learning regressors and classifiers to our problem and note the responses. First, we will split our dataset into train set and test set by the ratio of 74:26. Here, MSE stands for Mean Squared Error and R Square for R^2score. [2]

A. Linear Regression

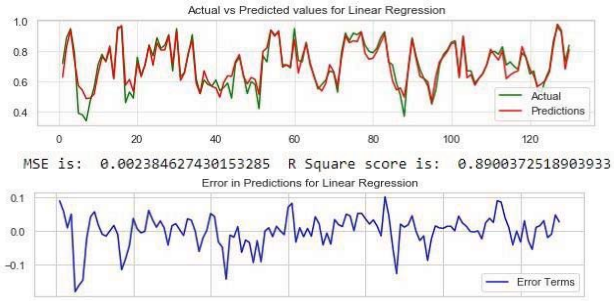


Fig 10. Predictions and Error of Linear Regression

B. Support Vector Regression

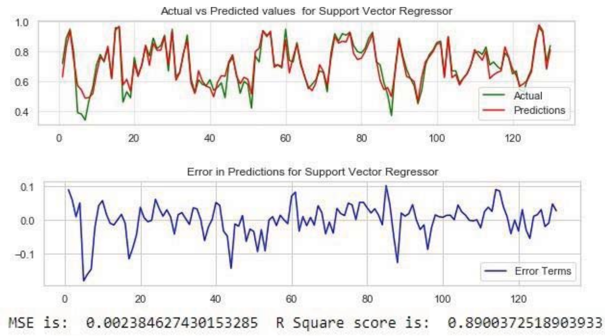


Fig 11. Predictions and Error of Support Vector Regression

C. Ridge Regression

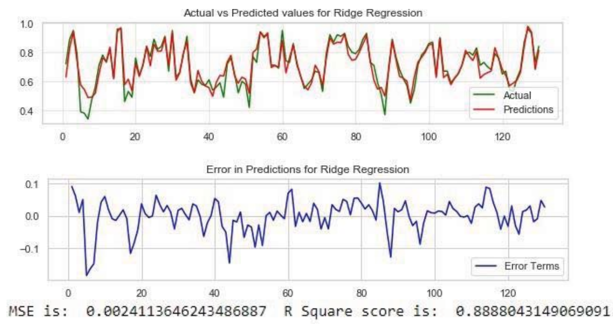


Fig. 12: Predictions and Error of Ridge Regression

D. Bayesian Ridge Regression

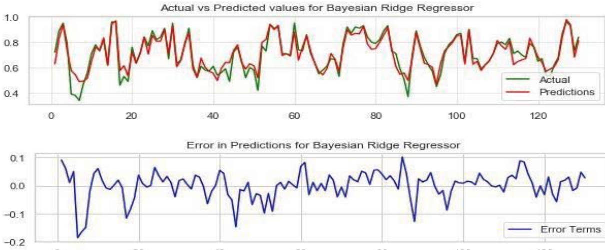


Fig.13: Predictions and Error of Bayesian Ridge Regression

E. Artificial Neural Networks

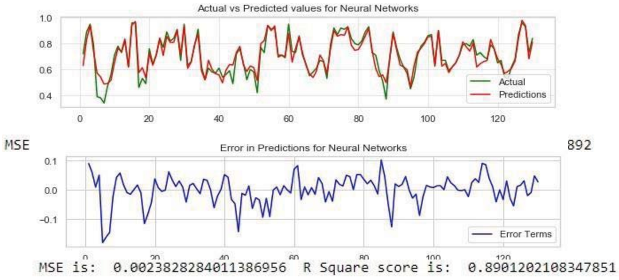


Fig.14: Predictions and Error of Artificial Neural Network

F. Random Forests Regression [6]

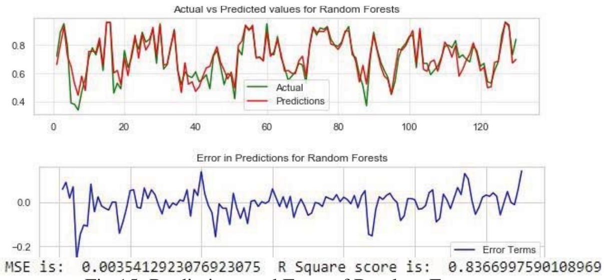


Fig 15. Predictions and Error of Random Forests

G. Ada Boost Regression

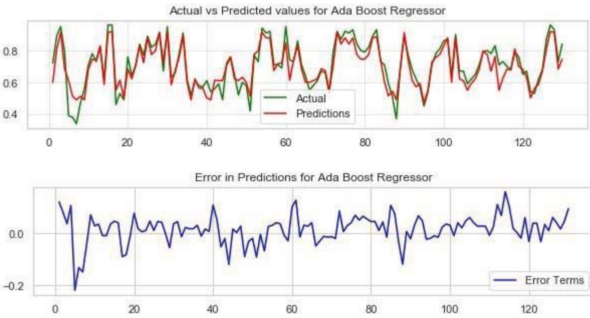


Fig 16. Predictions and Error of Ada Boost Regression

H. K-Nearest Neighbors

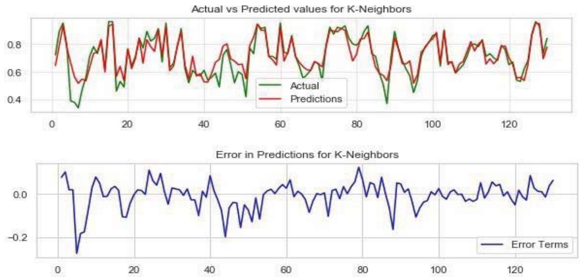
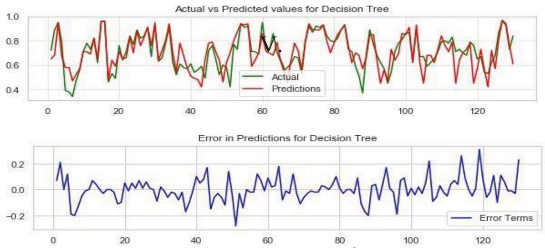


Fig 17. Predictions and Error of K-Nearest Neighbors Regression

H. Decision Trees




```
MSE is:  0.008666923076923076  R Square score is:  0.600340637224085
```

Fig 18. Predictions and Error of Decision Tree Regression

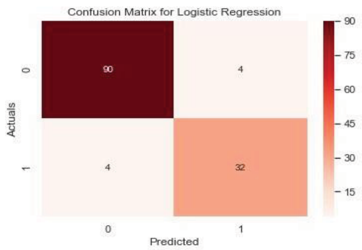
V. APPLYING MACHINE LEARNING CLASSIFIERS AND OBSERVING RESPONSE

Here, we will first change our data set labels from continuous values to discrete values (0 or 1: Admit or Not) by this algorithm:

new_y_train=[1 if x>0.8 else 0 for x in y_train] …..(2)
new_y_test=[1 if x>0.8 else 0 for x in y_test …..(3)

We selected 0.8 as our threshold here because it was giving the best values of precision and recall for the below algorithms. Plotting heat map of confusion matrix and calculating metrics by the following classification approaches. [5]

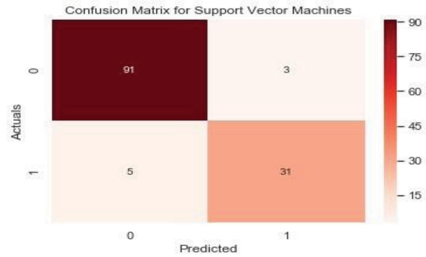
A. Logistic Regression




```
Accuracy score is: 0.9384615384615385
Precision score is:0.8888888888888888
Recall Score is:0.8888888888888888
F1 Score is: 0.8888888888888888
```

Fig 19. Confusion Matrix and metrics for Logistic Regression

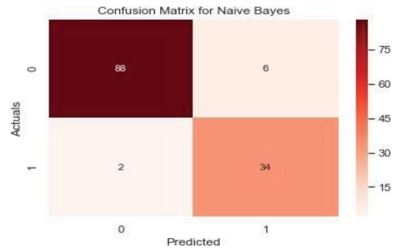
B. Support Vector Machine [7]




```
Accuracy score is: 0.9384615384615385
Precision score is:0.9117647058823529
Recall Score is:0.8611111111111112
F1 Score is: 0.8857142857142858
```

Fig 20. Confusion Matrix and metrics for Support Vector Machines

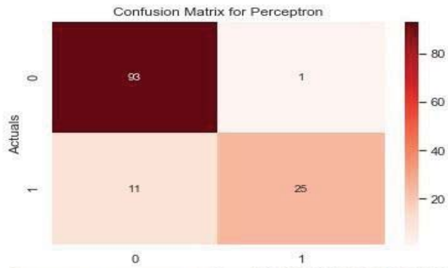
C. Naive Bayes Classifier




```
Accuracy score is: 0.9384615384615385
Precision score is:0.85
Recall Score is:0.9444444444444444
F1 Score is: 0.8947368421052632
```

Fig 21. Confusion Matrix and metrics for Naive Bayes Classifier

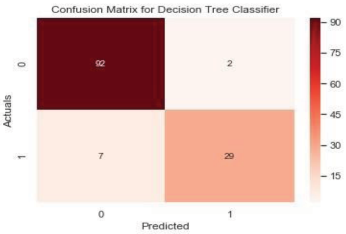
D. Perceptron




```
Accuracy score is: 0.9076923076923077
Precision score is:0.9615384615384616
Recall Score is:0.6944444444444444
F1 Score is: 0.8064516129032258
```

Fig 22. Confusion Matrix and metrics for Perceptron

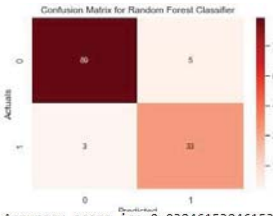
E. Decision Tree Classifier




```
Accuracy score is: 0.9307692307692308
Precision score is:0.9354838709677419
Recall Score is:0.8055555555555556
F1 Score is: 0.8656716417910448
```

Fig 23. Confusion Matrix and metrics for Decision Tree Classifier

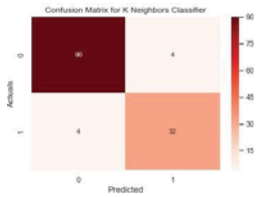
F. Random Forest Classifier [6]



```
Accuracy score is: 0.9384615384615385
Precision score is:0.868421052631579
Recall Score is:0.9166666666666666
F1 Score is: 0.8918918918918918
```

Fig 24. Confusion Matrix and metrics for Random Forest Classifier

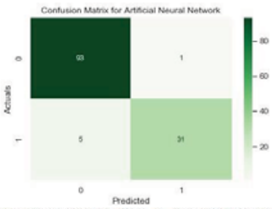
G. K-Nearest Neighbors Classifier



```
Accuracy score is: 0.9384615384615385
Precision score is:0.8888888888888888
Recall Score is:0.8888888888888888
F1 Score is: 0.8888888888888888
```

Fig 25. Confusion Matrix and metrics for K-Nearest Neighbors Classifier

H. Artificial Neural Network



```
Accuracy score is: 0.9538461538461539
Precision score is:0.96875
Recall Score is:0.8611111111111112
F1 Score is: 0.911764705882353
```

Fig 26. Confusion Matrix and metrics for Artificial Neural Network

VI. CLUSTERIZING APPLICANTS

Now, we will form clusters of similar applicants and label them if they are a correct fit for admission. [8] Elbow Method will be used to discover the best number of centroids.

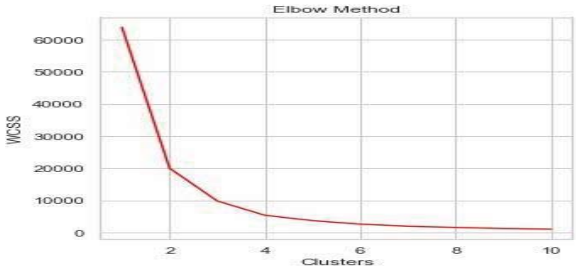


Fig 27. Plot of Least Sum of Squares vs Number of Clusters (Elbow Method)

We will take four clusters according to Elbow method. Plotting and labeling the clusters which are essentially the group of similar applicants pooled together based on CGPA and GRE Score. The labels are also shown in the graph

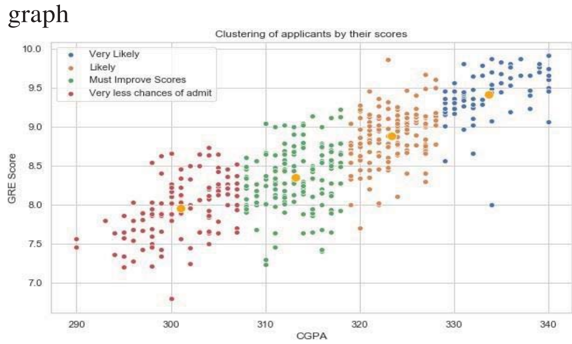
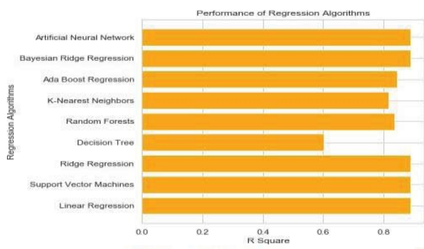


Fig 28. Formation of clusters with centroids and legends

VII. COMPARISON OF MACHINE LEARNING TECHNIQUES

For finding the best model to fit to this dataset, we compare all the algorithms with their metrics

A. Regression Algorithms



	Models	R Square Score
8	Artificial Neural Network	0.890120
0	Linear Regression	0.890037
1	Support Vector Machines	0.890037
2	Ridge Regression	0.888804
7	Bayesian Ridge Regression	0.888365
6	Ada Boost Regression	0.844897
4	Random Forests	0.836700
5	K-Nearest Neighbors	0.815416
3	Decision Tree	0.600341

	Models	Confidence Score
3	Decision Tree	1.000000
4	Random Forests	0.936137
6	Ada Boost Regression	0.792055
5	K-Nearest Neighbors	0.785162
0	Linear Regression	0.783450
1	Support Vector Machines	0.783450
8	Artificial Neural Network	0.783449
2	Ridge Regression	0.783389
7	Bayesian Ridge Regression	0.783344

Fig 29. Performance of Regression Algorithms with respect to R^2 Score and Confidence score

Analysis shows that, our Artificial Neural Network is the best fit here (Accuracy =0.89) and Decision tree overfits the data more according to the confidence score, that is 1.0 [3] [5].

B. Classification Algorithms

	Models	Accuracy
7	Artificial Neural Network	0.953846
0	Logistic Regression	0.938462
1	Support Vector Machines	0.938462
2	Naive Bayes	0.938462
5	Random Forest Classifier	0.938462
6	K-Nearest Neighbors Classifier	0.938462
4	Decision Tree Classifier	0.930769
3	Perceptron	0.907692

	Models	F1 Score
7	Artificial Neural Network	0.911765
2	Naive Bayes	0.894737
5	Random Forest Classifier	0.891892
0	Logistic Regression	0.888889
6	K-Nearest Neighbors Classifier	0.888889
1	Support Vector Machines	0.885714
4	Decision Tree Classifier	0.865672
3	Perceptron	0.806452

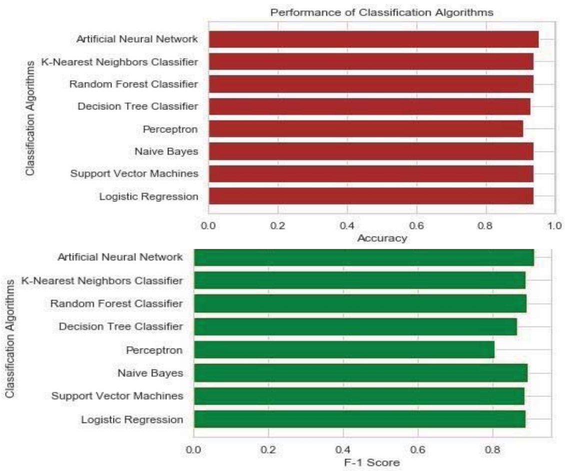


Fig 30. Comparison of Classification Algorithms based on Accuracy and F-Score

VII. CONCLUSION

Among our classification algorithms, Artificial Neural Network again performs best the based on Accuracy and F-Score (0.95 and 0.91). Even though it does not guarantee to converge to the global minima, it certainly does find a good optimum based on some random initialization of weights [3]. The work has yielded good results and are in accordance with the approach and performance has been satisfactory.

REFERENCES

[1] Raghunathan, Karthik. "Demystifying the American graduate admissions process." Web. 2010.

[2] Gupta, Narender, Aman Sawhney, and Dan Roth. "Will I get in? modeling the graduate admission process for American universities." 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). IEEE, 2016.

[3] Caruana, Rich, and Alexandru Niculescu-Mizil. "An empirical comparison of supervised learning algorithms." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

[4] Mohan Acharya, Asfia Armaan, Aneeta Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, International Conference on Computational Intelligence in Data Science 2019

[5] Géron, Aurélien. Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. " O'Reilly Media, Inc.", 2017.

[6] Dietterich, Thomas G. "Ensemble methods in machine learning." International workshop on multiple classifier systems. Springer, Berlin, Heidelberg, 2000.

[7] Schölkopf, Bernhard, Alexander J. Smola, and Francis Bach. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2002.

[8] Russell, Stuart J., and Peter Norvig. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited, 2016.