# Graduate Admission Predictions (May 2019)

Manpreet Singh, Manmeet Singh Gill and, Inderpreet Singh

**Abstract**— United States is one of the most popular higher education destinations for Indian students. Over 47,000 F1 visas were issued by the United States embassy to Indian nationals (Kandlakunta and Dubbudu, 2019) . Of these visas, most went to students attending technical graduate schools. We analyze and try to predict the chances of admission to University of California at Los Angeles. In this research, we analyze the graduate admission dataset available from Kaggle. We analyze and visualize the different features of the dataset. We apply supervised learning techniques on the dataset. Supervised machine learning techniques such as decision tree classification, naïve bayes and logistic regression are applied to the dataset. Of all the supervised learning algorithms, logistic regression gives the best accuracy.

**Index Terms**— Graduate Admissions, Machine Learning, Classification, Logistic Regression, Gaussian Naïve Bayes, Decision Tree Classifier, K-Means

## 1 Introduction

Education can be dated back to about 3400 B.C. when act of writing started (Nissen, 2016). Act of Ed-ucation can be defined as providing social, moral and intellectual information to an individual and involves the act of learning by that individual to gain beliefs, knowledge, skills and values. Education has proved to be one of the most important human necessities in today's modern world as it facilitates growth and ad-vancement in knowledge, living standards and other aspects as a global human civilization.

Education has been devided into various tiers. Ac-cording to education system of the United States of America, education can be divided and classified into various school systems. An individual has an option to either start from a Pre-school or a Primary school. After completing primary school, one can get admit-ted to middle school. After that, high school has to be completed before entering a college or a universi-ty where one can get Associates or Bachelor's degree respectively. After getting undergraduate (Bachelor) degree, one can get into graduate schools to get masters(Post Grad) degree and doctrate(doctoral) degree, which is considered to be highest degree in education system. An individual can also get into vocational schools for small courses for specializations or into professional schools to get professional degrees in fields like medicine, engineering, law etc.

Graduate schools which provide masters or doc-toral degrees provide admissions to students who have completed their undergrad degree on basis of their background in education.

Background of a student, applying for admission into a graduate school, can be further quantified into various factors like grade point average (GPA), Graduate Record Examination (GRE) scores etc.

In this project, we have used machine learning algorithms to predict the chances of admission of a stu-dent into a graduate school on basis of various fac-tors like GRE scores, university/school rankings, undergraduate GPA etc.

Section 2 of this paper provides a short insight into background and some previous work done into this topic by other researchers. Section 3 contains over-view of out research about different machine learning algorithms and explanations about why we chose certain algorithm and approach. Section 4 contains exploration and preprocessing of the data set. It pro-vides detailed insights into various features of the data set. Section 5 contains the information about the scoring metrics used to score the algorithms. Section 6 provides the results and findings of the project. In section 7, w compare the results and acuuracy of the algorithms while section 8 contains the conclusions about the project and research.

## 2 Background

In this section, we consider how the machine learning techniques can be applied to our graduate studies problem. Machine learning algorithms have been proven to automate the process of predicting something. Here, we provide some background of various machine learning approaches

### 2.1 Prediction Techniques/Strategies

Machine learning can be broadly divided into Supervised learning, learning in which y variable is known and data is labeled; Unsupervised learning, learning in which data is not labeled and y variable is not explicitly known; Reinforcement learning, learning in which data is dynamically changing and involves self learning.

As in our problem, the dataset is clearly labeled. Chances of admission can be clearly taken as y varia-ble. So, for such a problem, supervised learning is best suited.

### 2.2 Related Work

Some researchers have done some related work in this domain.

A researcher from Mercedes Bens. Dr. Binu (Binu, 2017) has used Artificial Neural Networks (ANN) to predict the chances of admission into the university. He also added second hidden layer to the model after performing feature scaling. He was able to achieve 90% accuracy using ANN.

Dr. Kru Ceng, a researcher from Gazi University in Turkey has tried to use kernelized support vector machines (SVM) to solve the same problem (Ceng, 2018). SVM works by building a hyperplane in the data points to do classification. He was able to attain about 89.9% accuracy using SVM. He also used deci-sion tree to do classification. He was only able to get 83% accuracy using decision trees.

## 3 ML MODELS/ALGORITHMS

Our dataset has dependent variable y (Chance of Admit) which is continuous variable ranging from 0 to 1. It indicates the student's chances of admission based on the dataset features such as GRE score, TOEFEL score etc. Since our dataset is labelled, it makes it feasible to apply supervised learning algorithms to train the model for the predictions. There are two main techniques that we can follow in supervised learning to train the model for the predictions.

a. Classification: - It is task of classifying elements of given set into multiple groups/labels. It is type of supervised learning where the labels are predefined, model categorize the new probabilistic oservations into predefined categories.

b. Regression: - Regression aims to model the relationship b/w dataset fatures and dependent variable which is continuous in nature. The trained model will predict the continuous numerical value. For e.g a given candidate has 0.7 chances of admission.

Since our aim for training model is to predict whether the candidate will have higher chances of admission into graduate program or not, we have used classification as our basis for the predictions. Moreover, whether the student will get admission or not can be interpreted or converted to labels of 1 or 0 respectively. Label 1 will indicate that student has very high chances of admission, whereas label 0 indictes that given student has low chances of admission. In order to assign these labels 1 or 0, we need to set some threshold value which can indicate that students with Chance of Admit value more than or equal to threshold value will have very high chances of admission, hence will receive label 1. On the other hand, students who have Chance of Admit value lower than threshold value will have lower chances of admission, hence will receive label 0. The threshold value we have selected is 0.80.

## 4 DATASET FEATURES

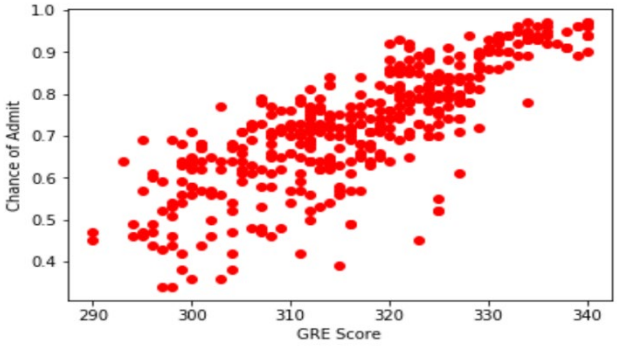
We have obtained our graduate admission dataset from Kaggle. It has following features along with their range: -

- GRE Scores (290 to 340)
- TOEFL Scores (92 to 120)
- University Rating (1 to 5)
- Statement of Purpose (1 to 5)
- Letter of Recommendation Strength (1 to 5)
- Undergraduate CGPA (6.8 to 9.92)
- Research Experience (0 or 1)
- Chance Of Admit (0 to 1)

Most of the features are intuitive to understand. The feature "University Rating" is the university the student has graduated from. Since this dataset is mainly fetched from the collection of statistics of Indian students who are trying to get admission in US universities, the Indian univerisities are broken into different tier/category universities, hence the name of the feature as "University Rating". Now we need to know the trends of the features. How features relate with the chance of admission variable. How features affect the chance of admission of students.

### 4.1 GRE SCORE

GRE is the most common graduate admission test wshich is required by all US universities for graduate programs




admission. Like SAT & CAT, GRE assess the candidates's analytical, crtitcal writing and quantative reasoning skills. We will see how GRE score affects the chance of admit variable.

Figure – 1

As we can clearly see in the above figure that as GRE scores increases, the chance of admit increases highly. This shows clearly that GRE score is important deciding feature in our prediction model.

### 4.2 TOEFL SCORE

International applicants are required to submit the TOEFL scores when applying for graduates' programs in US universities. Following is the trend of TOEFL score feature with respect to chance of admit.

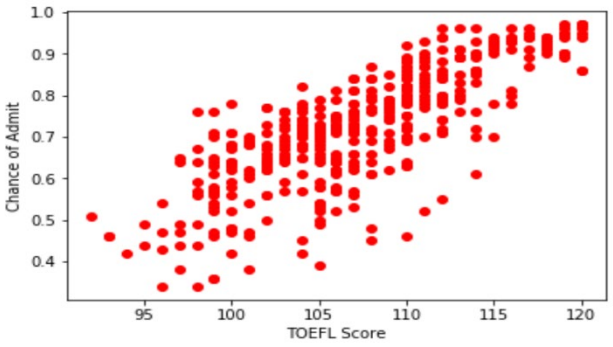



Figure – 2

Higher TOEFL scores corresponds to higher chance of admit. It can be concluded that TOEFL score will be good feature in our training model.

### 4.3 UNIVERSITY RATING

The university here refers to the applicant's undergraduate university. It is fair to conclude that every university does have same reputition. Few universities are recognized best, few are reputed as average. This may or may not have impact on the chance of admission, but it is paremeter that can play good role in deciding which student to be admitted when the two candidates (with different universities) have same profile.

For example, in India, universities are categorized into different categories/tiers.

**TIER 1**- IITs Bombay, Kanpur, KGP, Delhi, Madras, Guwahati, BHU, Hyderabad and Roorkee. In addition to these IITs: BITS pilani pilani campus, IIIT Hyd, and ISM Dhanbad. These will be considered rating 5 universities in our case.

**Tier 2** -NEW IITs like Mandi, Gandhinagar, Patna, Bhubaneswar etc. DAIICT, BIT Mesra, IIIT Bangalore, IIIT Gwalior, IIIT Kancheepuram, IIIT Jabalpur. Rest of the NITs. These may be considered rating 4 universities in our case.

Now, we can see or analyse how university rating affects chance of admit.

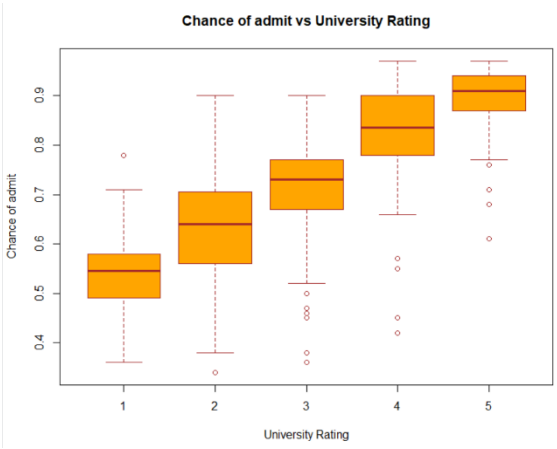



Figure – 3

We can see from the above box plot that universities with high rating follow increasing trend as the mediam value in each box plot group increases from 1 to 5.

### 4.4 SOP

The statement of purpose serves as a method to intoduce applicant's skills ad interests to the program applicant is applying to. The dataset has SOP featue whose values range from 1 to 5. This means that applicant's SOP have marked or rated with the values from low (0) to high (5). Again, SOP may or may not be good feature in deciding but it may come into role when two applicant's profile matches exactly, then SOP can be deciding factor.

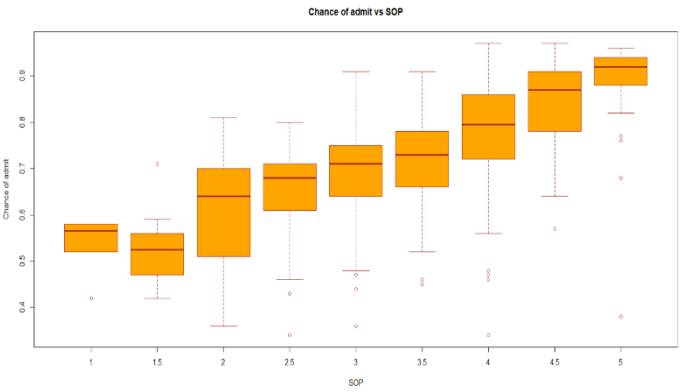



Figure - 4

As we can see in Figure-4 box plot (SOP vs Chance of Admit), the mediam values for each box plo group increases, showing increasing trend for the chance of admit.

### 4.5 LOR

Letter of Recommendations play significant role in US. They show applicant's work ethics, behavior, relations with professors/empoyers. Let's see how LOR plays role in deciding the chance of admit in our problem domain.

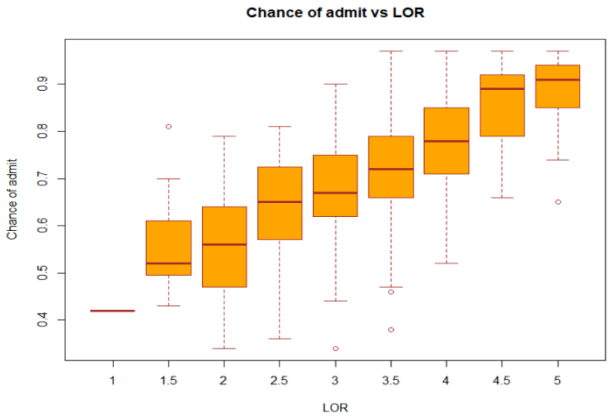



Figure – 5

The median value in Boxplot for each group increases from 1 to 5 showing increasing trend w.r.t chance of admit.

### 4.6 CGPA

In India, the CGPA (GPA in US) has range of 0 – 10. We can observe how CGPA affects chance of admit in our problem.

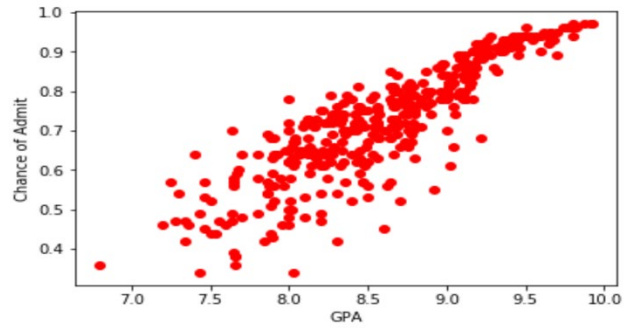



Figure – 6

It is clearly visible that CGPA will paly significant role as deciding factor, hence CGPA is important feature.

### 4.7 RESEARCH EXPERIENCE

We have total 219 candidates with research experience and 181 candidates with no research experience.

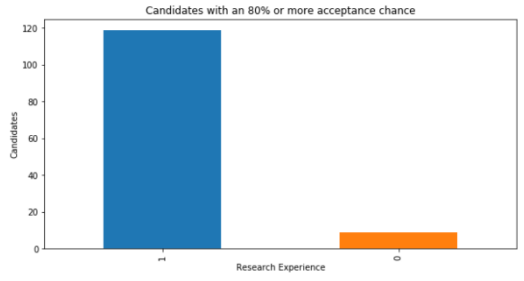



Figure –

As we can see, dataset has 119 candidates with research experience who have equal to or more than 80% chances of admission, whereas, there are 9 candidates which do not have research experience, but still have 80% or more chances of admission.

## 5 SCORING METRICS USED

We have used following metrics to compare the accuracy of our trained model by applying different algorithms.

a. Precsion_Score: - TP/ (TP + FP)

b. Recall_Socre: - TP/(TP + FN)

c. F1_Score: - 2/ ((1/precision) + (1/recall))

TP: - True Positives, FP: - False Positives
FN: - False Negatives

**Precision** Score: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. **Precision is a good measure to determine, when the costs of False Positive is high.** In email spam detection, a false positive means that an email that is non-spam (actual negative) has been identified as spam (predicted spam). The email user might lose important emails if the precision is not high for the spam detection model.

Recall Score : Recall is the ratio of correctly predicted positive observations to the all postive observations. **Recall shall be the model metric when there is a high cost associated with False Negative.** In fraud detection or sick patient detection. If a fraudulent transaction (Actual Positive) is predicted as non-fraudulent (Predicted Negative), the consequence can be very bad for the bank.

F1 Score: F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. **F1 Score might be a better measure to use if we need to seek a balance between Precision and Recall AND there is an uneven class distribution (large number of Actual Negatives, which is our case).**

**Actual Positives: 117, Actual Negatives : 283 [Total = 400 observations]**

## 6 RESULTS & FINDINGS

**Logistic Regression**: - We used binary logistic model to do classification for assigning labels 1 or 0 to the test data observations. The threshold value for label 1 is more than or equal to 0.80 value of chance of admit. Y values less than 0.80 will assing label 0.
Following are the results for test observations with logistic regression moel.

Precision_Score: 0.9583333333333334

Recall_Score: 0.7931034482758621

F1_Score: 0.8679245283018867

**Gaussian Naïve Bayes**: - It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Following are the test results achieved by applying Gaussain Naïve Bayes for our classification problem.

Precision_Score: 0.9333333333333333

Recall_Score: 0.9655172413793104

F1_Score:0.9491525423728815

**Decision Tree Classifier**: - A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a student has chance of admit or not), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.
Precision_Score: 0.9333333333333333

Recall_Score: 0.9655172413793104

F1_Score:0.9491525423728815

**K Nearest Neigbors (KNN)**:- Knn is non-parametric method used for classification and regression. In KNN classification, the output is class membership. The membership/label is assigned by plurality of neighbor votes. Knn algorithm requires input of K closest neighbors from the training set. These K neighbors decide what membership will be assigned to the test object based on their votes.

Precision_Score: 0.9285714285714286

Recall_Score: 0.896551724137931

F1_Score: 0.912280701754386

**Confusion Matrix of Logistic Regression Model**: -
According to Confusion Matrix, the logistic regression model predicted that 23 candidate's Chances of Admit are greater than 80%. In reality, 22 of them have a Chance of Admit greater than 80%. In total, 29 candidate's Chances of Admit are greater than 80%.

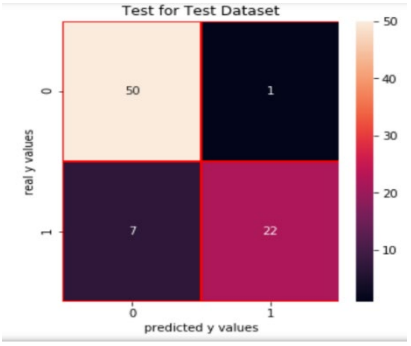



Figure - 7

Similarly, according to Confusion Matrix, the model predicted that 57 candidate's Chances of Admit are less than or equal to 80%. In reality, 50 of them have a Chance of Admit less than or equal to 80%. In total, 51 candidate's Chances of Admit are less than or equal to 80%.

Following are the confusion matrix for the Gaussian Naïve Bayes and the Decision Tree Classifier.

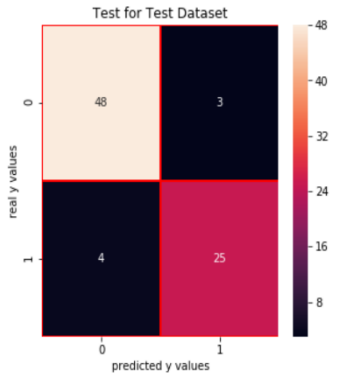



Figure –8 – Gaussian CM

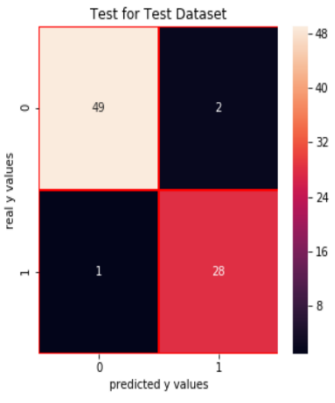



Figure-9-DecisionTree CM

## 7 COMPARISON OF ACCURACY

In the below figure, we have plotted the F1 scores of three algorithms that we used for predictions.

**F1 Scores: -**

Log. Regression: - 0.86792452830188

Gaussian NB: - 0.9491525423728815

Decision Tree: - 0.867924528301886

K Neighbors: - 0.912280701754386

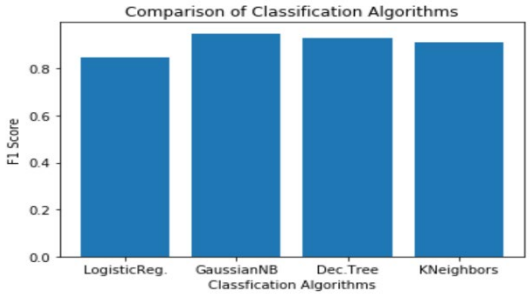



Figure – 10

As graph for F1 scores predicts that Gaussian Naïve Bayes is the most successful algorithm for our predictions with 94 percent F1 score which is comparitvely very high as compared to other two algorithm's F1 scores.

## 8 Conclusion

Using real world data from UCLA, we showed that supervised machine learning algorithms are able to pre-dict admission chances with a fairly accurate rate. We analyzed the dataset and applied both supervised and unsupervised machine learning algorithms. We applied logistic regression, gaussian naïve Bayes, decision trees classifier and K-NN algorithms on the dataset. Whereas all the algorithms predict with a fair accuracy, logistic regression seems to be the best candidate. We also con-didered whether regression or classification would be suitable for this problem. We applied unsupervised machine learning algorithm ( K means ) to further ana-lyze the relation between datapoints. We also tried to show that some of the features like CGPA, TOEFL scores and GRE scores show have much clear and linear rela-tion with chances of admis-sion. Whereas, other features like LOR and SOP are much more scatter and can't be related directly, even though chances do increase slightly with increase in these fea tures. These features are more scattered than others in the dataset.

## 9 GitHUB LINK

https://github.com/manpreetsjsu/Predict_Graduate_Admissions

## Acknowledgment

We would also like to show our gratitude to Dr. Vishnu Pendyala, Data Scientist at Cisco and Professor at San Jose State University for sharing his knowledge of Machine Learning and Big Data with us during the course of this research, and for providing us the opportunity for this research project.

## References

[1] Binu, R. (2017). University admission in era of Nano De-grees | Kaggle. [online] Kaggle.com. Available at: https://www.kaggle.com/biphili/university-admission-in-era-of-nano-degrees [Accessed 2 May 2019].

[2] Ceng, K. (2018). Analyzing the Graduate Admission EDA & ML | Kaggle. [online] Kaggle.com. Available at: https://www.kaggle.com/kralmachine/analyzing-the-graduate-admission-eda-ml [Accessed 2 May 2019].

[3] Nissen, H. (2016). The Evolution of Writing | Denise Schmandt-Besserat. [online] Sites.utexas.edu. Available at: https://sites.utexas.edu/dsb/tokens/the-evolution-of-writing/ [Accessed 2 May 2019].

[4] Kandlakunta, A. and Dubbudu, R. (2019). Trump effect is real. Student Visas to Indians down 40% in 2 years. [online] FACTLY. Available at: https://factly.in/students-visas-us-16-2016-total-non-immigrant-visas-us-increase/ [Accessed 2 May 2019].R. Ni-cole, "The Last Word on Decision Theory," J. Computer Vision, submitted for publication. (Pending publication)

## Biography

**Singh, Manpreet** Manmeet is a senior undergraduate student pursuing Software Engineering from SJSU. He has past experience as web intern at SJSU H&A Marketting Department.

**Gill, Manmeet Singh** is a senior undergraduate student pursuing Software Engineering from SJSU. Being a fourth-year student, he is also a member of Tau Beta Pi, scholar society for Engineering. He has past experience as an android developer in DisplayRide.

**Singh, Inderpreet** is currently a third-year software engineering major at San Jose State University. He has past experience working as a research assistant where he helped in analyzing and solving a real-world problem through machine learning