**PROJECT REPORT ON**

**“Graduate Admission Prediction”**

**(Using ML with Python)**

Submitted By:

**1) PrajaktaKeer**

 **2) SakshiDeore**

**Of Pune Institute of Computer Technology, Pune.**

**Sr. No 27, Pune-Satara Road, Dhankawadi,**

**Pune-411 043.**

Under the Guidance of:

**Mr. Gurvansh Singh (M.Tech)**

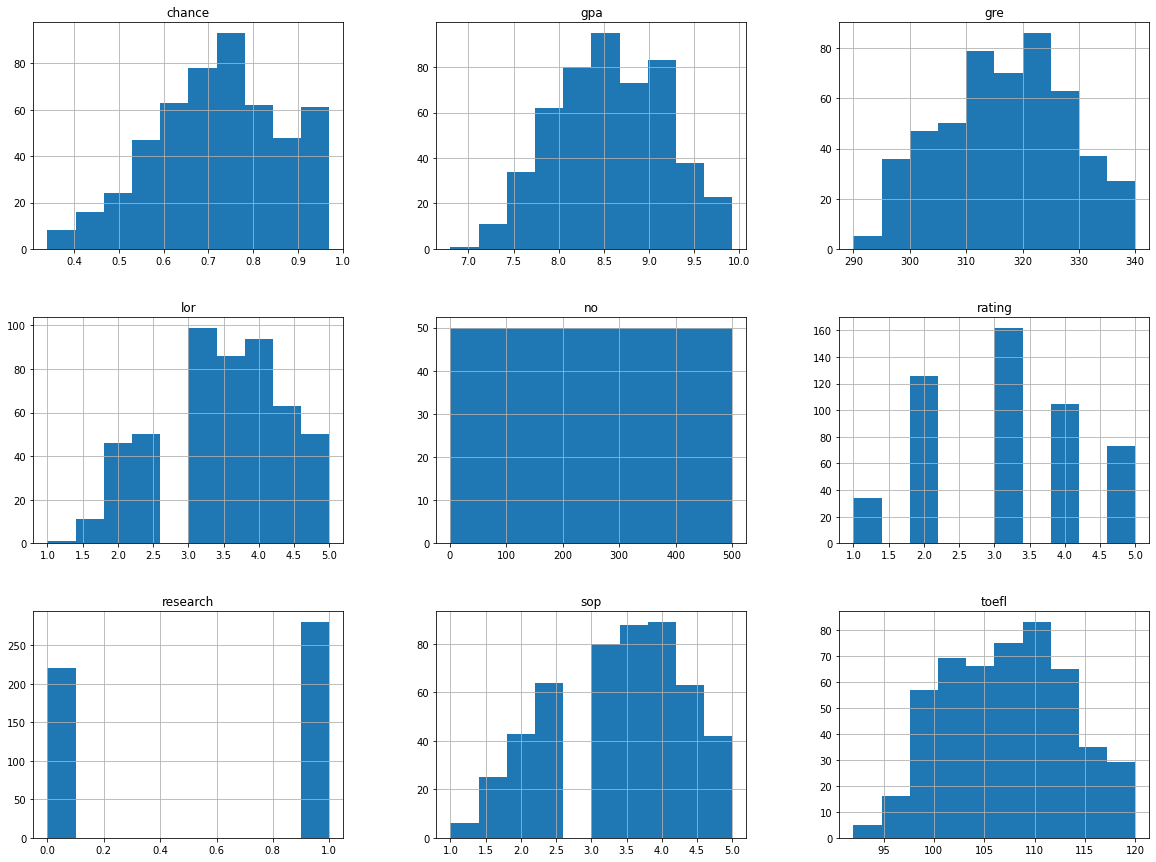
**Knowledge Solutions India.**

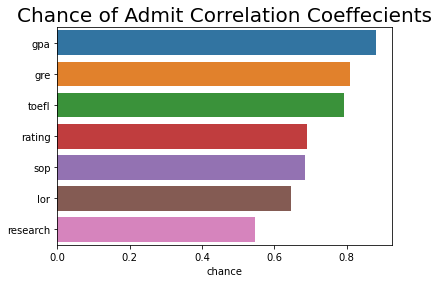
**Table of Content**

|  |  |  |
| --- | --- | --- |
| **SR NO.** | **TOPIC** | **PAGE NO** |
| 1 | Table of figures/graphs | 3-4 |
| 2 | Abstract | 5 |
| 3 | Introduction | 6-7 |
| 4 | Software libraries used | 8-9 |
| 5 | Mathematics | 10 |
| 6 | Conclusion | 11 |
|  |  |  |

**Table Of Figures/Graphs:**

**Histogram of Numeric Variables:**

****

****

**ABSTRACT**

This project tries to understand Graduate Admissions process by specifically analyzing different features of around 500 student like GRE score, Toefl score etc. Chance of admission of students for postgraduate programhas been given data set. After extensive data pre-processing, we have tried to build admission prediction model. The key to analyzing Graduate Admissions data is to analyze data in buckets rather than considering all in one bucket. The project aims to help students choose the right Universities by predicting whether a student will be admitted to a specific University. Similarly, this model can be used by Graduate Admission Committee to predict the admission of particular student in the University.

**Introduction**

This project analyzes various features such as GRE Score, TOEFL Score, CGPA, University Rating, Research, Rating of Universities of various students who had applied to get admission in foreign universities and predicts the chance of student getting the admission.

The dataset provided was pretty small. We shuffled the data to avoid patterns. By performing exploratory analysis on the data we found that data was quite normal as it did not contain any null values or outliers. The next step was data pre-processing. Feature scaling helped to normalize the features and to bring them at a same scale. Then we extracted the features. We then split the dataset into test set and train set at a ratio of 25 : 75.

We developed 4 models for this dataset:

1) Multiple Linear Regression(MLR)

2) MLR with Principal Component Analysis(PCA)

3) Random Forest Regressor(RFR)

4) RFR with PCA

**MLR:**

Later we trained the Linear Regression model to predict the chance of admission using the test data set. We then calculated the errors(MSE, RMSE, R2 score). Then we performed backward elimination to remove less significant features. Again by calculating the errors we achieved an accuracy of 83.9%. Then we created a scatter plot of actual values vs. predicted values.

**MLR with PCA:**

We used PCA to eliminate features and by taking n\_components as 1 we reduced the number of features from 7 to 1. We trained the Linear Regression model to predict the chance of admission using the test data set. By doing so, we achieved an accuracy of 79.5%. We plotted a graph of x vs. the chance of admit and also a scatter plot of actual values vs. predicted values.

**RFR:**

We used RFR model to predict the chance of admission of students in postgraduate program. In this model, first we have found data table for all features giving mean, maximum, minimum values and some other values. Then we grouped data according to the university rating. From this, we came to know that chance of admission of students with higher university rating is high. Then we splitted our dataset into train and test datasets. After that feature scaling is done. Then by doing some observation we put output as 1 for students having chance>0.82 while we put 0 for those having chance<0.82. Then we trained our random forest regressor model. By doing so, we have achieved an accuracy of 85.3%.

**RFR with PCA:**

In this model, first we did feature elimination using PCA. Then we trained our RF model. From this model we achieved the accuracy of 78.5%.

**Software Libraries Used**

We created all 4 models using Python.

**Python -** Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace.

**Python libraries used -**

**Numpy**:NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

**Pandas**: In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license.

* read\_csv() function to read the dataset

**Matplotlib**: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

* to create scatter plot
* to create 2D plot

**Sklearn** :Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines.

* from pre-processing module imported StandardScalarclass to perform feature scaling
* from decomposition module imported PCA to reduce the number of features
* from model\_selection module imported train\_test\_split() function to split the dataset
* from linear\_model imported LinearRegression class to implement MLR
* imported the metrics module to calculate errors

**Mathematics**

## Solving linear regression using Ordinary Least Squares – general formula

A simple linear regression function can be written as:

yi=β0+β1xi+ϵi, i=1,2...nyi=β0+β1xi+ϵi, i=1,2...n

We can obtain n equations for n examples:

y1=β0+β1x1+ϵ1y2=β0+β1x2+ϵ2⋮yn=β0+β1xn+ϵny1=β0+β1x1+ϵ1y2=β0+β1x2+ϵ2⋮yn=β0+β1xn+ϵn

If we add n equations together, we get:

∑yi=nβ0+β1∑xi+∑ϵi∑yi=nβ0+β1∑xi+∑ϵi

Because for linear regression, the sum of the residuals is zero. We get:

∑yi=nβ0+β1∑xi                 (1)∑yi=nβ0+β1∑xi                 (1)

If we use the Ordinary Least Squares method, which aims to minimize the sum of the squared residuals. We define C to be the sum of the squared residuals:

C=(β0+β1x1−y1)2+(β0+β1x2−y2)2+...+(β0+β1xn−yn)2C=(β0+β1x1−y1)2+(β0+β1x2−y2)2+...+(β0+β1xn−yn)2

This is a quadratic polynomial problem. To minimize C, we take the partial derivatives with respect to β1β1 and set the results to 0 and we get:

∑xiyi=∑xiβ0+β1∑x2i     (2)∑xiyi=∑xiβ0+β1∑xi2     (2)

Solving equations (1) and (2), we can then get:

β0=(∑x2i)(∑yi)−(∑xi)(∑xiyi)n∑x2i−(∑xi)2β1=n(∑xiyi)−(∑xi)(∑yi)n∑x2i−(∑xi)2β0=(∑xi2)(∑yi)−(∑xi)(∑xiyi)n∑xi2−(∑xi)2β1=n(∑xiyi)−(∑xi)(∑yi)n∑xi2−(∑xi)2

**CONCLUSION**

Introduction:

Machine learning is a sub-domain of computer science which evolved from the study of

pattern recognition in data, and also from the computational learning theory in artificial

intelligence. It is the first-class ticket to most interesting careers in data analytics today[1]. As

data sources proliferate along with the computing power to process them, going straight to the

data is one of the most straightforward ways to quickly gain insights and make predictions.

Machine Learning can be thought of as the study of a list of sub-problems, viz: decision

making, clustering, classification, forecasting, deep-learning, inductive logic programming,

support vector machines, reinforcement learning, similarity and metric learning, genetic

algorithms, sparse dictionary learning, etc. Supervised learning, or classification is the machine

learning task of inferring a function from a labeled data [2]. In Supervised learning, we have a

training set, and a test set. The training and test set consists of a set of examples consisting of

input and output vectors, and the goal of the supervised learning algorithm is to infer a function

that maps the input vector to the output vector with minimal error. In an optimal scenario, a

model trained on a set of examples will classify an unseen example in a correct fashion, which

requires the model to generalize from the training set in a reasonable way. In layman’s terms,

supervised learning can be termed as the process of concept learning, where a brain is exposed to

a set of inputs and result vectors and the brain learns the concept that relates said inputs to

outputs. A wide array of supervised machine learning algorithms are available to the machine

learning enthusiast, for example Neural Networks, Decision Trees, Support Vector Machines,

Random Forest, Naïve Bayes Classifier, Bayes Net, Majority Classifier[4,7,8,9] etc.,

In this project, we have predicted the chance of admission of a studentin universities for master’s program. From the study, it has been observed that CGPA and GRE score have more impact on chance of Admission.

Accuracy of the 4 models is as follows:

1)MLR - 83.93%

2)MLR with PCA - 78.53%

3) RFR – 85.3%

4) RFR with PCA –78.5%

From the above figures it is clear that Random Forest Regressor is the best model with an accuracy of 85.3%