



KNOWLEDGE SOLUTIONS INDIA

Project report

On

University Admission Prediction

Submitted by

Dheeraj

Sravani

Usha Sree

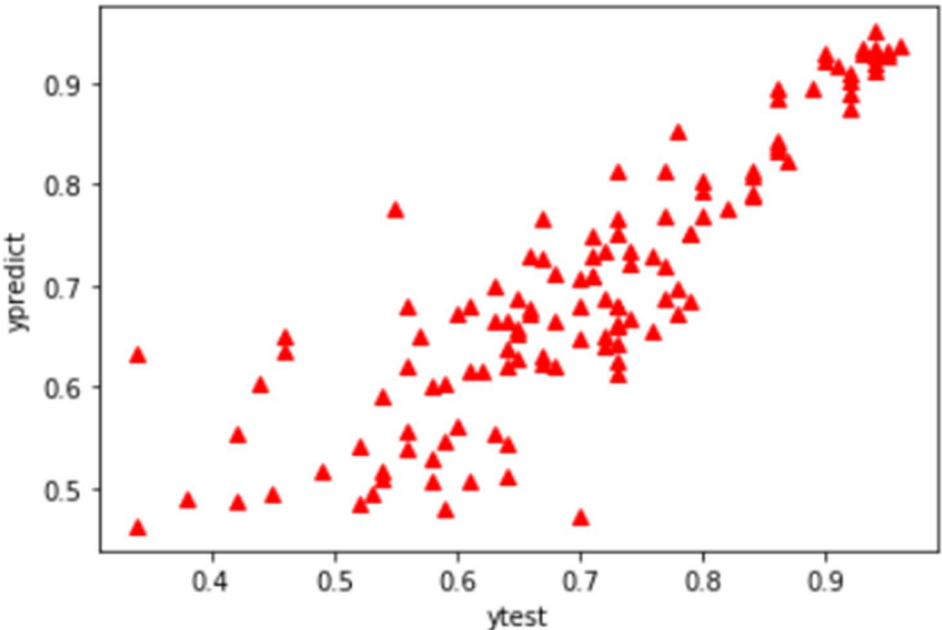
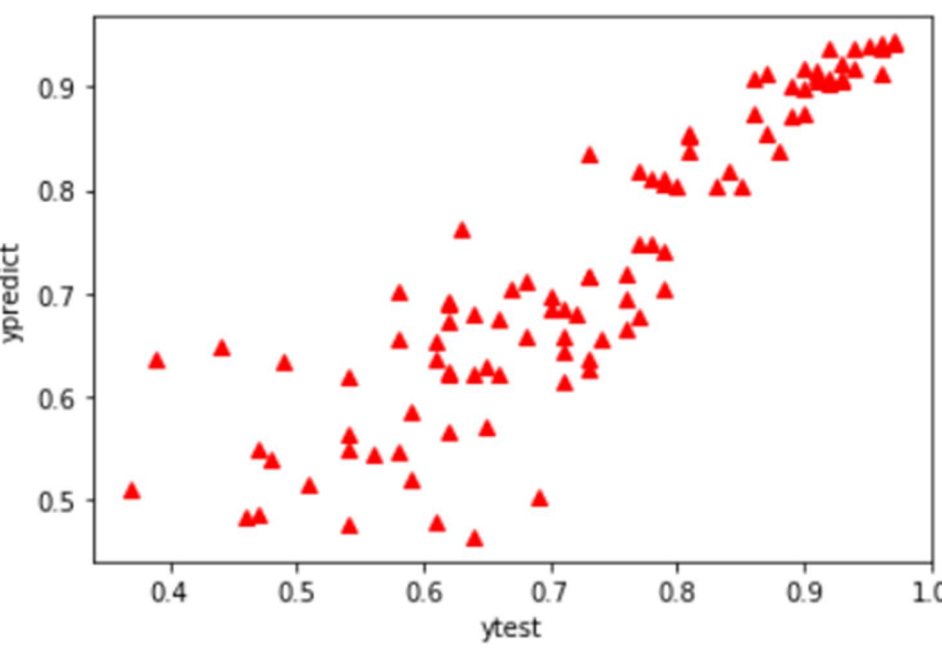
Jasmin



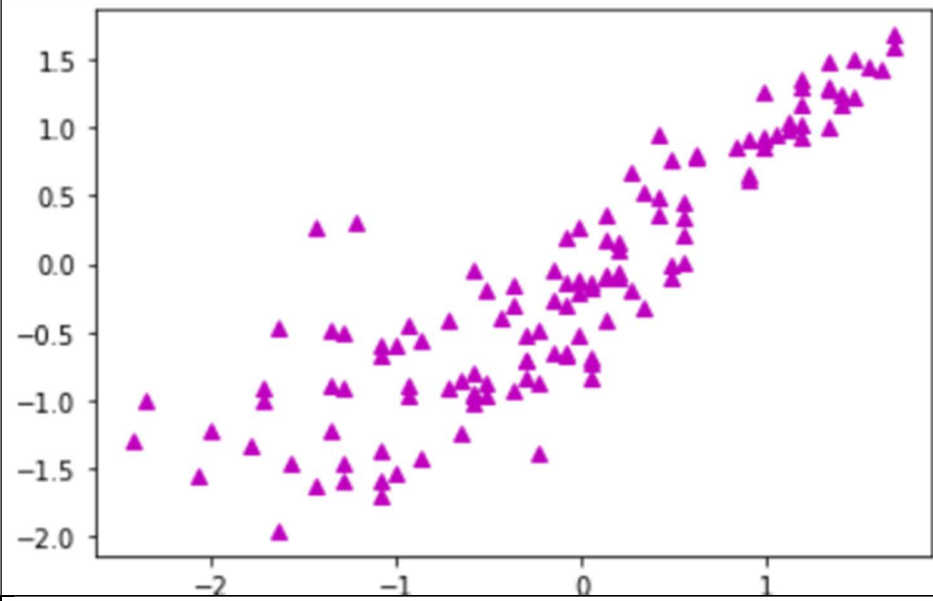
Table of contents

sno	topic	Page no
1	Table of figures	1
2	abstract	3
3	introduction	4
4	Libraries used	5
5	Algorithms used	5
6	Code/code snippets	6
7	conclusion	7

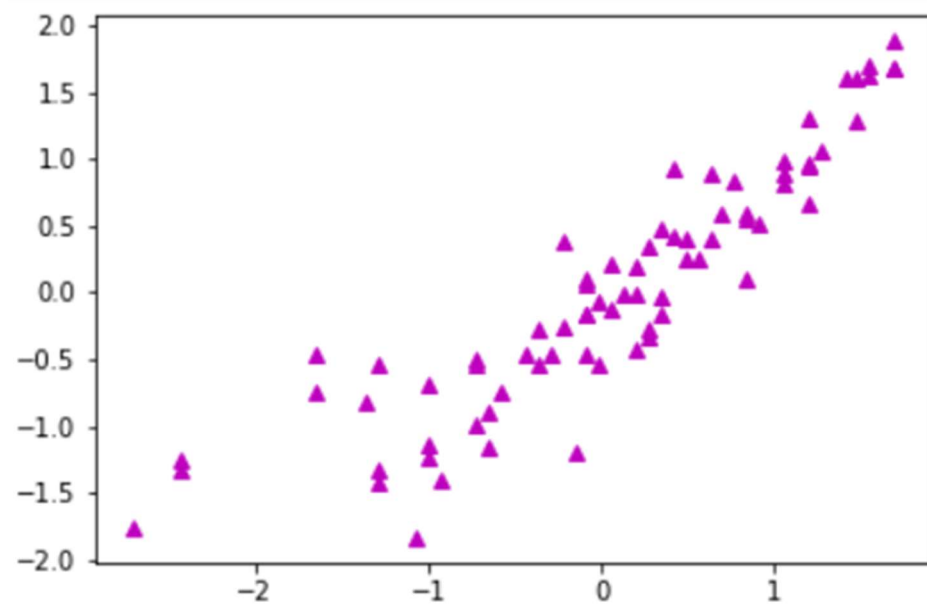
Table of graphs

Model/graph name	graph
Random forest(y test vs y pred)	 <p>A scatter plot showing the relationship between predicted values (ypredict) on the y-axis and test values (ytest) on the x-axis for a Random Forest model. The data points are represented by red triangles. The plot shows a strong positive correlation, with most points falling along the diagonal line where ytest equals ypredict. The x-axis ranges from approximately 0.35 to 0.95, and the y-axis ranges from approximately 0.45 to 0.95. There is a slight spread of points around the diagonal, particularly in the middle range.</p>
Random forest with pca (ytest vs ypred)	 <p>A scatter plot showing the relationship between predicted values (ypredict) on the y-axis and test values (ytest) on the x-axis for a Random Forest model with PCA. The data points are represented by red triangles. The plot shows a strong positive correlation, with most points falling along the diagonal line where ytest equals ypredict. The x-axis ranges from approximately 0.35 to 1.0, and the y-axis ranges from approximately 0.45 to 1.0. The points are more tightly clustered along the diagonal compared to the first plot, indicating better predictive performance.</p>

MLr with pca



mlr



Abstract:in the todays era we can see so many students who want to pursue their higher studies in other countries. Some of the main countries where students show interest to go and study are United States of America and china. Year by year the number of students applying and studying in these foreign universities are increasing drastically. Majority of the surveys say that compared to other countries Indian students are in majority to apply for these foreign universities. Due to the number of students increased in applying for these universities each applicant has to face a tough competition to get admission in their dream university. Generally students in India seek the help of consultancies who then analyse their their profile and tell which university they would get.in this procedure the consultancies charge a heavy amount of fees and also all consultancy services are not reliable in India. Apart from these there are many websites or bloggers who analyse the students profile and suggest which university would fit for them. The drawback of the currently available resources are very limited, some charge very high fee and some are not truly reliable.to overcome this problem we have come up with a project named “University Admission predictor” which is developed using different machine learning algorithms, useful for students whose dream is to study abroad can check their chance of admission percentage in the desired college. Based on the students profile details such as gre, toefl,research etc the model can predict the chance of admit.

Introduction: a person's education plays an important role in their life. While planning for education, students generally have many questions in their mind regarding courses, universities, expenses involved etc. Securing admission in their dream university is on their main concern. It is often seen that students want to pursue their higher education from universities which have global recognition. Majority of the Indian students' first preference is the United States of America where world's highly reputed universities, wide range of courses etc. are available. According to the Department of Education, United States, there are about 2500 private universities in the country.

The majority of the students who apply for these universities are from India and China. In the past decade, India has seen a huge increase in the number of students opting to pursue their education from foreign universities like in the USA. In India, students are finding it difficult to get admission in highly ranked colleges. India is one of the leading countries which produce software engineers every year. Due to this, the competition to achieve a job in elite companies has become tough. Due to this reason, many students motivate themselves to pursue higher education like PG in their respective domains.

Majority of the international universities follow similar guidelines for providing admission to students. Universities take into consideration different factors like score based on aptitude based examinations like General Record examination (GRE), command over their English language is judged based on their score in English competency test like Test of English as Foreign Language (TOEFL), their letter of recommendation, research papers they have published etc. Based on the overall profile of student, the university admission team decides whether to admit or reject the student.

Every candidate has to take all the required examinations and build a strong profile to secure admission in their dream universities. Once the candidates have made their profile, they apply for the universities where they aim to secure admission. The students have to shortlist the universities which are best known for courses they are looking for and also students must have an idea of their chance of admission in that particular university. This particular task of shortlisting the universities based on high chance of admit is very difficult and many students end up in applying for many universities. There are several portals, consultancies who analyze a student's profile and tell the chance of admission, other universities etc. but they charge very high fee and are very limited in countries like India. Moreover, some of the consultancies in India are not reliable due to which students finally end up in trouble.

The main objective of our project is to create machine learning models that predict the chance of admit which will help many students. Multiple machine learning algorithms are used to develop this like multiple linear regression, random forest, principal component analysis along with random forest, principal component analysis with multiple linear regression.

This project finally helps students saving extra amount of time and money they have to spend at educational consultancy firms. And also helps students to limit the number of applications to a small number.

Software libraries used: The libraries used are numpy, matplotlib, pandas, seaborn, sklearn

Numpy: Numpy which stands for numerical python is a library consisting of multi dimensional array objects and a collection of routines for processing those arrays using numpy mathematical and logical operations on arrays can be performed

Matplotlib: matplotlib is a visualisation library in python. it is a comprehensive library for creating static, animating visualisations in python

Pandas: pandas library is one of the fast and easy to use data analysis and manipulation tool built on python programming language. it is a free and open source library.

Seaborn: seaborn is a library in python that uses matplotlib underneath to plot graphs. it helps in visualizing random distributions. it provides a high level interface for drawing attractive and informative statistical graphs.

Sklearn: sklearn is an open source machine learning library for python. it features various algorithms like support vector machines, random forests, k-neighbours, linear regression etc

Algorithms used

Multiple Linear Regression: MLR also known as multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. the goal of multiple linear regression is to model the linear relationship between independent variables and dependent variables

Random Forest: random forest is a supervised machine learning algorithm. the forest it builds is an ensemble of decision trees usually trained with bagging method. the general idea of bagging method is that a combination of learning models increases the overall result.

Principal component analysis: pca also known as principal component analysis is a technique to bring out strong patterns in a dataset by suppressing variations. it is mainly used to clean data sets and analyse. this algorithm is based on few mathematical ideas namely variance and co variance

Code/Code snippets

Importing libraries

Importing the Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Checking for any outliers or null values

```
# checking the null values
df.isnull().sum()
```

```
GRE Score      0
TOEFL Score    0
University Rating  0
SOP            0
LOR            0
CGPA           0
Research       0
Chance of Admit  0
dtype: int64
```

Multiple Linear Regression:

train and evaluate Multiple Linear Regression models

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, accuracy_score
```

```
LinearRegression_model = LinearRegression()
LinearRegression_model.fit(X_train, y_train)
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

Random Forest

train and evaluate random forest models

```
from sklearn.ensemble import RandomForestRegressor
RandomForest_model = RandomForestRegressor(n_estimators=100, random_state=0)
RandomForest_model.fit(X_train, y_train)
```

Principal component analysis

pca

```
from sklearn.decomposition import PCA
pca = PCA(n_components=2)
X = pca.fit_transform(X)
print(pca.explained_variance_ratio_)
```

```
[0.67519343 0.10596446]
```


The different kpis calculated

```
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
from math import sqrt
k = X_test.shape[1]
n = len(X_test)
RMSE = float(format(np.sqrt(mean_squared_error(y_test_orig, y_predict_orig)), '.3f'))
MSE = mean_squared_error(y_test_orig, y_predict_orig)
MAE = mean_absolute_error(y_test_orig, y_predict_orig)
r2 = r2_score(y_test_orig, y_predict_orig)
adj_r2 = 1-(1-r2)*(n-1)/(n-k-1)

print('RMSE =', RMSE, '\nMSE =', MSE, '\nMAE =', MAE, '\nR2 =', r2, '\nAdjusted R2 =', adj_r2)

RMSE = 0.067
MSE = 0.0044448219999999999
MAE = 0.047431999999999999
R2 = 0.8050237221094269
Adjusted R2 = 0.8010035926683841
```

Conclusion: we have used four kinds of algorithms namely multiple linear regression, random forest, pca with mlr, pca with random forest to find the chance of admit. The accuracy of the models also varies random forest having an accuracy of 75.3%, random forest with pca having an accuracy of 80.5%, mlr with pca having an accuracy of 76.3%, mlr with 82.5%. so the conclusion can be made that using using Multiple Linear Regression we have obtained the highest accuracy which is **82.5%**

