Admission Prediction

Internship report submitted in fulfillment of the requirements for the Degree of

B.Tech

in

CSE

By

Deepak Sati

ENROLLMENT NO.

18103011

Department of Computer Science Engineering and Information Technology
JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY
(Declared Deemed to be University U/S 3 of UGC Act)
A-10, SECTOR-62, NOIDA, INDIA
August - 2021

Table of Contents

Prof. (Dr.) Vikas Saxena, SMIEEE
Ph.D. (CSE), M.E. (CE), B.Tech. (CS&IT)
Head-Department of CSE&IT
Jaypee Institute of Information Technology
A-10, Sector-62, Noida, India, 201307
+91-120-2594273, vikas.saxena@jiit.ac.in

E-Certificate Number: CSE\SIT-2021\005

To whomsoever it may concern

This is to certify that Mr. Deepak Sati (Enrollment no. 18103011) a student of B. Tech. in Computer Science & Engineering from Jaypee Institute of Information Technology has successfully completed the summer training under the supervision of Dr. K Vimal Kumar from 21st June 2021 to 3rd August 2021 (6 weeks).

We wish him all success in the future.

Vikas Saxena

Project Objective

To predict the admission of candidate in university using various exam scores and 6 different models in machine learning

1. SVR : (Support Vector Regression)

SVM or Support Vector Machine is **a linear model for classification and regression problems**. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

2. Random Forest Classifier

A random forest is a machine learning technique that's **used to solve regression and classification problems**. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees.

3. Logistic Regression with max_iter=1000

Logistic regression is **a supervised learning classification algorithm used to predict the probability of a target variable**. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. Mathematically, a logistic regression model predicts P(Y=1) as a function of X.

4. KNN : (K-Nearest-Neighbors)

The k-nearest neighbors (KNN) algorithm is a **simple, supervised machine learning algorithm** that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slower as the size of that data in use grows.

5. CLF : Decision Tree Classifier

It is a **tree-structured classifier**, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node.

6. Linear Regression

Linear Regression is a **machine learning algorithm based on supervised learning**. ... Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output).

Now experimenting with 5 different model

1. SVR : (Support Vector Regression)
2. Random Forest Classifier
3. Logistic Regression with max_iter=1000
4. KNN : (K-Nearest-Neighbors)
5. CLF : Decision Tree Classifier

In [359]:
```
# importing SVM model from scikit

from sklearn.svm import SVR

from sklearn.tree import DecisionTreeClassifier

from sklearn.linear_model import LogisticRegression

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import explained_variance_score

model_SVR =  SVR(kernel = 'rbf')

model_CLF  = DecisionTreeClassifier(criterion = "gini",random_state = 100,max_depth=3, min_samples_leaf=5)
```

Figure 1

Experimentation with 5 different ML models

Project Description

Student admission problem is very important in educational institutions. This paper addresses machine learning models to predict the chance of a student to be admitted to a master's program. This will assist students to know in advance if they have a chance to get accepted. The machine learning models are multiple SVR, linear regression, k-nearest neighbor, random forest,Logistic regression and CLF. Experiments show that the Linear regression model surpasses other models.Young workers who want to stand out in their jobs are always looking for higher degrees that can help them in improving their skills and knowledge. As a result, the number of students applying for graduate studies has increased in the last decade. The dataset presented in this paper is related to the educational domain.

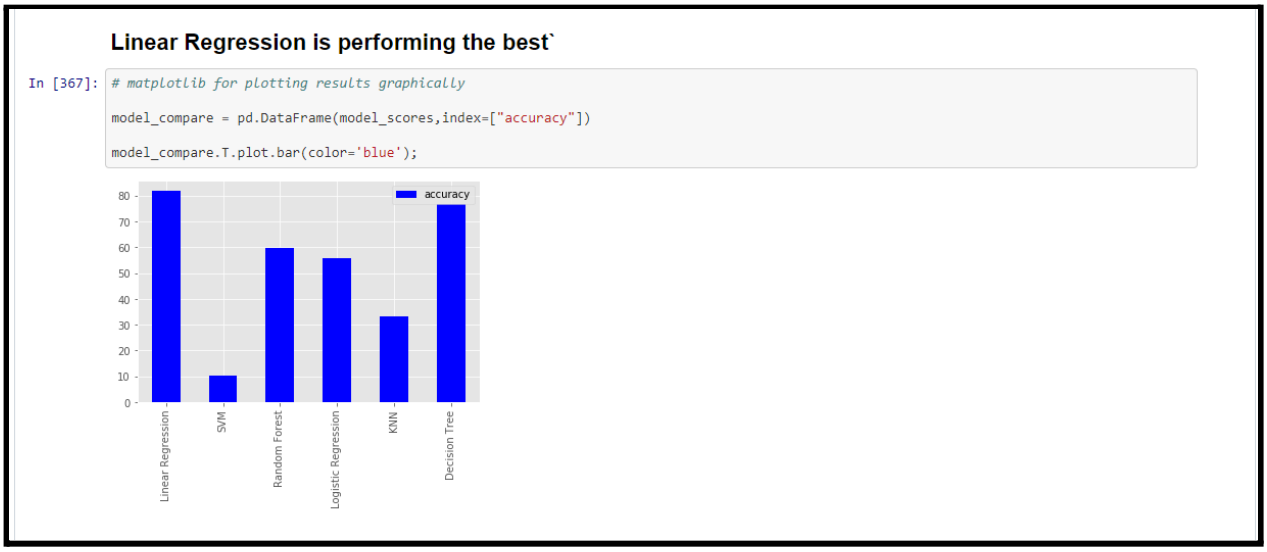

Figure 2

Linear regression giving highest accuracy

Admission is a dataset with 500 rows that contains 7 different independent variables which are: •Graduate Record Exam1 (GRE) score. The score will be out of 340 points. • Test of English as a Foreigner Language2 (TOEFL) score, which will be out of 120 points. • University Rating (Uni.Rating) that indicates the Bachelor University ranking among the other universities. The score will be out of 5 1 Graduate Record Examination www.ets.org/gre 2 Test Of English as

Foreign Language www.ets.org/toefl

	Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
1	1	337	118	4	4.5	4.5	9.65	1	0.92
2	2	324	107	4	4	4.5	8.87	1	0.76
3	3		104	3	3	3.5	8	1	0.72
4	4	322	110	3	3.5	2.5	8.67	1	0.8
5	5	314	103	2	2	3	8.21	0	0.65
6	6	330	115	5	4.5	3	9.34	1	0.9
7	7	321	109		3	4	8.2	1	0.75
8	8	308	101	2	3	4	7.9	0	0.68
9	9	302	102	1	2	1.5	8	0	0.5
10	10	323	108	3	3.5	3	8.6	0	0.45
11	11	325	106	3	3.5	4	8.4	1	0.52
12	12	327	111	4	4	4.5	9	1	0.84
13	13		112	4	4	4.5	9.1	1	0.78
14	14	307	109	3	4	3	8	1	0.62
15	15	311	104	3	3.5	2	8.2	1	0.61
16	16	314	105	3	3.5	2.5	8.3	0	0.54
17	17	317	107	3	4	3	8.7	0	0.66
18	18	319	106	3	4	3	8	1	0.65
19	19	318	110	3	4	3	8.8	0	0.63
20	20	303	102	3	3.5	3	8.5	0	0.62

Figure 3

Data set with 500 different records

• Statement of purpose (SOP) which is a document written to show the candidate's life, ambition and the motivations for the chosen degree/ university. The score will be out of 5 points. • Letter of Recommendation Strength (LOR) which verifies the candidate's professional experience, builds credibility, boosts confidence and ensures your competency. The score is out of 5 points • Undergraduate GPA (CGPA) out of 10 • Research Experience that can support the application, such as publishing research papers in conferences, working as research assistant with university professors (either 0 or 1). One dependent variable can be predicted which is the chance of admission, that is according to the input given will be ranging from 0 to 1.ML Model uses above data set for predicting the percentage chance % of admission of candidate in a university using his exam scores.Model was saved in a pickle file for future testing without running the whole model again and again

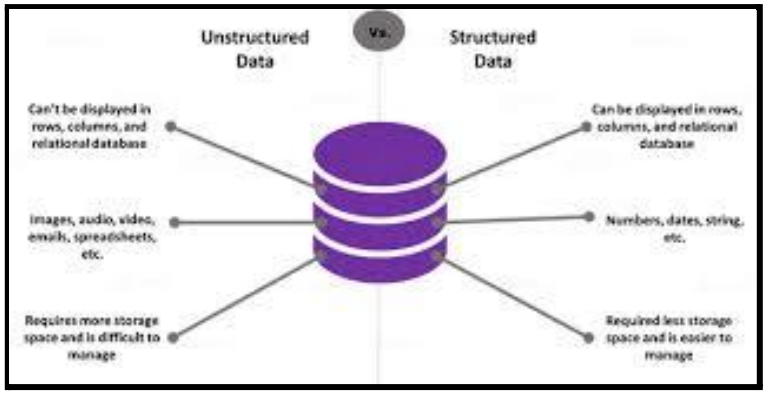

Figure 4

Our data set falling under structured category

Contribution

1. 6 Weeks per day work report can be found here

https://docs.google.com/spreadsheets/d/1OAaHA9aQMj2G7w1DlNpdCsAanuQlKwd7XbHtjPibx1Q/edit?usp=sharing

2. Learning and exploring new topics and tools in machine learning.

3. Research on the topic for an in-house internship which provides some applications.

4. Responsible for the whole user interface both frontend and backend.

5. Research on the model which will suit best and give great accuracy

6. Experimentation with 5 different machine learning models

7. Research on errors and how to resolve them effectively

8. Improving and styling the final layout

9. Preparing per day work report and final report to be submitted to college

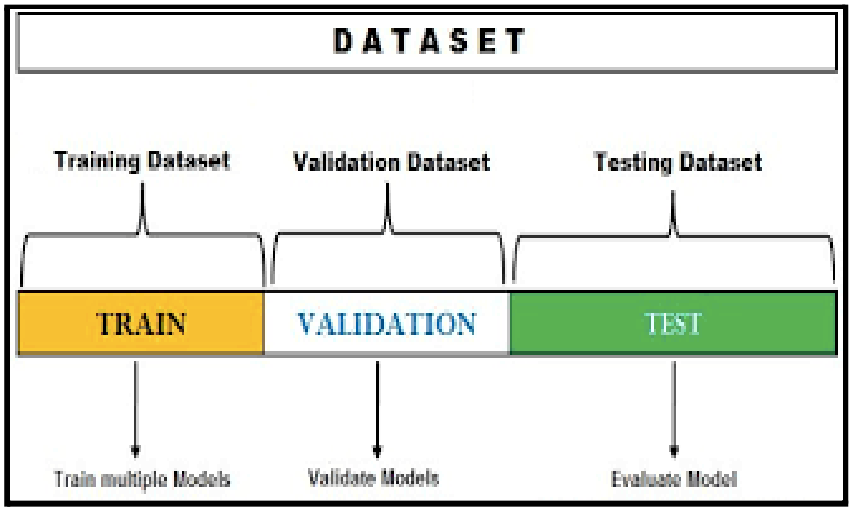

Figure 5

Breaking our prediction process into 3 parts

Skills/technologies developed

Anaconda , Jupyter notebook, google colab, pandas, matplotlib, numpy, scikit, python, kaggle

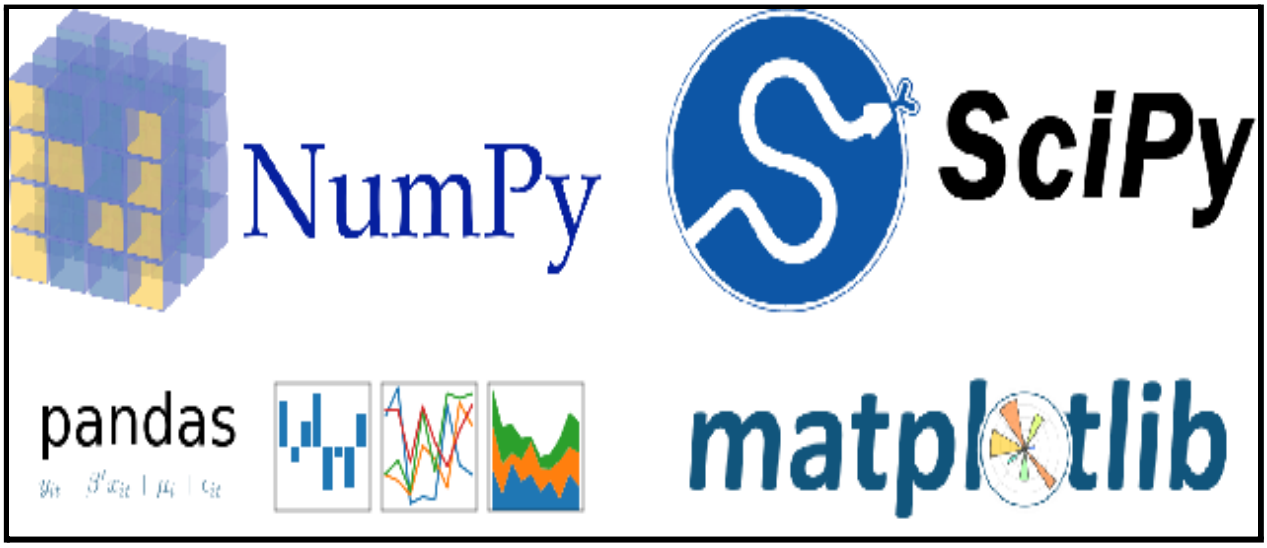

Figure 6

Various testing methodologies in structured machine learning

Conclusion

In this paper, machine learning models were performed to predict the opportunity of a student to get admitted to a master's program. The machine learning models included are multiple linear regression, k-nearest neighbor, random forest, and SVM.

Experiments show that the Linear regression model surpasses other models with an accuracy of 81.91 %.As for the future work, more models can be conducted on more datasets to learn the model that gives the best performance.

References

1. A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger and J. V. Guttag, "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration", J. Biomed. Inform., vol. 53, pp. 220-228, Feb. 2015

2. R. Krishna Reddy , " Big data analytics and computational Intelligence ", Chirala, Andhra Pradesh, India, pp. 44-63, 23-25 March 2017

3. Min Chen , "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities " , China 8869 - 8879 , 26 April 2017

4. P. Groves, B. Kayyali, D. Knott and S. van Kuiken, The Big Data Revolution in Healthcare: Accelerating Value and Innovation, 2016

5. B. Qian, X. Wang, N. Cao, H. Li and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction", Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070-1093, 2015