Machine Learning Project

Acknowledgement

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them.

I owe my deep gratitude to our project guide Ms. Mousita Dhar, Mr. Malay Mitra who took keen interest on my project work and guide me all along till the completion of my project work by providing all the necessary information for developing a great project.

I heartily thank our internal project guide, Dr. Chandan Banerjee, HOD of IT(NSEC) for his guidance and suggestions during the project work.

I am thankful and fortunate enough to get constant encouragement, support and guidance from all teaching support and guidance from all teaching staffs of Webtek labs which helped me in successfully completing my project work. Also I would like to extend my sincere esteems to all staff in laboratory for their timely support.

Kaushik Sarkar
Sem- VII,Year- 4th
Information Technology

CERTIFICATE OF APPROVAL

The project "Prediction of University Admission based on GMAT, GPA score "made by Kaushik Sarkar is hereby approved as a creditable study for the BACHELOR OF TECHNOLOGY in INFORMATION TECHNOLOGY and presented in a manner of satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned this project only for the purpose for which it is submitted.

_____	_____
Mr. MalayMitra	Ms. Mousita Dhar
(project in-charge)	(project in-charge)

Candidate Declaration

I hereby declare that I have undertaken the Industrial Training at "WEBTEK LABS" during a period from 20th JANUARY to 14th FEBRUARY in partial fulfillment of requirements for the award of degree of B. TECH(INFORMATION TECHNOLOGY) at NETAJI SUBHASH ENGINEERING COLLEGE , KOLKATA. The work which is being present in the training report submitted to Department of INFORMATION TECHNOLOGY at NETAJI SUBHASH ENGINEERING COLLEGE, KOLKATA is an authentic record of training work.

Student Name **Signature of the student**

Kaushik Sarkar _____

The four weeks industrial training Viva-Voce Examination of _____

_____ has been held on _____ and accepted.

_____ _____

Signature of Internal Examiner

Signature of External Examiner

Contents

Introduction

1.1)Python: Python is a clear and powerful object-oriented programming language ,comparable to Perl, Ruby, Scheme, or java.

Python's features:

- Uses an elegant syntax ,making the programs we write easier to read.
- Is an easy-to-use language that makes it simple to get our program working .this makes python ideal for prototype development and other ad-hoc programming tasks.
- Python's interactive modes makes it easy to test sort snippets of code. There also a bundled development environment called idle.
- Is easily extended by adding new modules implemented in a compiled language such as c or c++.
- Can also be embedded into an application to provide a programmable interface .
- Runs anywhere, including Mac os ,Windows ,Linux and Unix, with unofficial builds also available for android and IOS.

Python's programming language features:

1.python supports object oriented programming with class and multiple inheritance

2.Code can be grouped in modules and packages

3.This language supports raising and catching exceptions, resulting in cleaner error handling

4.Python contain advance programming features such as generators and list comprehensions.

5.Python automatic memory management frees you from having to allocate and free memory in your code

Python's Versions:

1.Web Development
2.Data Analysis
3.Machine Learning
4.Internet of things
5.GUI development
6.Image Processing
7.Data Visualization
8.Game Development

1.2)ANACONDA :

The open-source Anaconda distribution is the easiest way to perform Python or R Data Science and machine learning on Linux ,Windows and Mac os with over 15 million user worldwide

Anaconda's Features: It is the industry standard for developing testing and training on a single machine,enabling individual data scientists to :

1.Quickly download 1500+python/R data science packages.
2.Analyze data scalability and performance with Dusk , Pandas, NumPy etc.
3.Manage dependencies, libraries and environment with conda.
4.Visualize result with Matplotlib, Bokeh, datashader and Holoviews.

Introduction to Machine Learning

Machine learning is an application of artificial intelligence(AI) that provides system the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observation or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

How machines learn:

Although a machine learning model may apply a mix of different techniques, the methods for learning can typically be categorized as three general types:

- **Supervised learning**: The learning algorithm is given labeled data and the desired output. For example, picture of dogs labeled "dog" will help the algorithm identify the rules to classify pictures of dogs.
- **Unsupervised learning**: The data given to the learning algorithm is unlabeled, and the algorithm is asked to identify patterns in the input data. For example, the recommendation system of an e-commerce website where the learning algorithm discovers similar items often bought together.
- **Reinforcement learning**: The algorithm interacts with a dynamic environment that provides feedback in items of rewards and punishments. For example, self-driving cars being rewarded to stay on the road.

Application:

1.Fraud Detection

2.Email Spam Filtering

3.Handwriting Recognition

4. Recommendation Engine

5.Medical Diagnosis

6. Employee Salary Prediction etc.

Introduction to NumPy and Pandas

NumPy:

- NumPy (Numeric Python) is Linear Algebra Library for Python.
- NumPy enriches the programming language Python with powerful data structures for efficient computation of multi-dimensional arrays and matrices.
- A NumPy array is a grid of values, all of the same type, and is indexed by a tuple of nonnegative integers.
- The number of dimensions is the rank of the array; the shape of an array is a tuple of integers giving the size of the array along each dimension.

Pandas:

- Pandas is the most popular python library that is used for data analysis
- It provides highly optimized performance with back-end source code is purely written in C or Python.
- We analysis data in pandas with
    1. Series(1-d array)
    2. DataFrame(2-d array)
- Using pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data -- load, prepare, manipulate, model and analyze.

Steps of Machine Learning

To apply the learning process to real-world tasks, we'll use a five-step process. Regardless of the task at hand, any machine learning algorithm can be deployed by following these steps:

1.**Data collection:** The data collection step involves gathering the learning material an algorithm will use to generate actionable knowledge. In most cases, the data will need to be combined into a single source like a text file, spreadsheet or database.

2. Data exploration and preparation: The quality of any machine learning project is based largely on the quality of its input data. Thus, it is important to learn more about the data and its nuances during a practice called data exploration. Additional work is required to prepare the data for the learning process. This involves fixing or cleaning so-called "messy" data, eliminating unnecessary data, and recoding the data to conform to the learner's expected inputs.

3. Model training: By the time the data has data has been prepared for analysis, you are likely to have a sense of what you are capable of learning from the data. The specific machine learning task chosen willrepresent the data in the form of a model.

4.Model evaluation: Because each machine leaning model results in a biased solution to the learning problem, it is important to evaluate how well the algorithm learns from its experience. Depending on the type of model used, you might be able to evaluate the accuracy of the model using a test dataset or you may need to develop measures of performance specific to the intended application.

5. Model improvement: If better performance is needed, it becomes necessary to utilize more advanced strategies to augment the performance of the model altogether. You may need to supplement your data with additional data or perform additional preparatory work as in step of this process.

Admission Prediction Based on GMAT, GPA Score

Aim: Our aim was to predict the chances of admission for student according to his/her GMAT score, GPA score, age & work experience.

Model used to analyze the project: Regression Model(Random Forest Regression).

Random Forest Regression: A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Pros and Cons of Random Forest

Pros

The following are the advantages of Random Forest algorithm –

- It overcomes the problem of overfitting by averaging or combining the results of different decision trees.
- Random forests work well for a large range of data items than a single decision tree does.
- Random forest has less variance then single decision tree.
- Random forests are very flexible and possess very high accuracy.
- Scaling of data does not require in random forest algorithm. It maintains good accuracy even after providing data without scaling.
- Random Forest algorithms maintains good accuracy even a large proportion of the data is missing.

Cons

The following are the disadvantages of Random Forest algorithm –

- Complexity is the main disadvantage of Random forest algorithms.
- Construction of Random forests are much harder and time-consuming than decision trees.
- More computational resources are required to implement Random Forest algorithm.
- It is less intuitive in case when we have a large collection of decision trees.
- The prediction process using random forests is very time-consuming in comparison with other algorithms.

Training Work Undertaken

- **Data collection:**
  Data set was provided by trainer but generally Kaggle is used to find and explore the data sets.
  Kaggle is an online community of data scientists and machine learners owned by Google LLC. Kaggle allows user to find and publish data sets explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers and enter competitions to solve data science challenges.
- **Dataset:** candidates.csv

  It is a comma separated value (CSV) file i.e. when we will open the file by any other editor like notepad the column will be separated by com.

	A	B	C	D	E	F
1	gmat	gpa	work_experience	age	admitted	
2	780	4	3	25	2	
3	750	3.9	4	28	2	
4	690	3.3	3	24	1	
5	710	3.7	5	27	2	
6	780	3.9	4	26	2	
7	730	3.7	6	31	2	
8	690	2.3	1	24	0	
9	720	3.3	4	25	2	
10	740	3.3	5	28	2	
11	690	1.7	1	23	0	
12	610	2.7	3	25	0	
13	690	3.7	5	27	2	
14	710	3.7	6	30	2	
15	680	3.3	4	28	1	
16	770	3.3	3	26	2	
17	610	3	1	23	0	
18	580	2.7	4	29	0	
19	650	3.7	6	31	1	
20	540	2.7	2	26	0	
21	590	2.3	3	26	0	
22	620	3.3	2	25	1	
23	600	2	1	24	0	
24	550	2.3	4	28	0	
25	550	2.7	1	23	0	
26	570	3	2	25	0	
27	670	3.3	6	29	1	
28	660	3.7	4	28	1	
29	580	2.3	2	26	0	
30	650	3.7	6	30	1	
31	760	3.3	5	30	2	
32	640	3	1	23	0	
33	620	2.7	2	24	0	
34	660	4	4	27	1	
35	660	3.3	6	29	1	
36	680	3.3	5	28	1	

- gmat= GMAT score of the candidate
- gpa = GPA score of the candidate
- work_experience= Year of work experience of the candidate
- age= age of the candidate
- admitted= No. of candidates got admission

Data visualization:

Data visualization is the discipline of trying to understand the data by placing it in a visual context. Python offers great graphing libraries such as **Matplotlib** and **seaborn** that come packed with lots of different features.

Matplotlib:

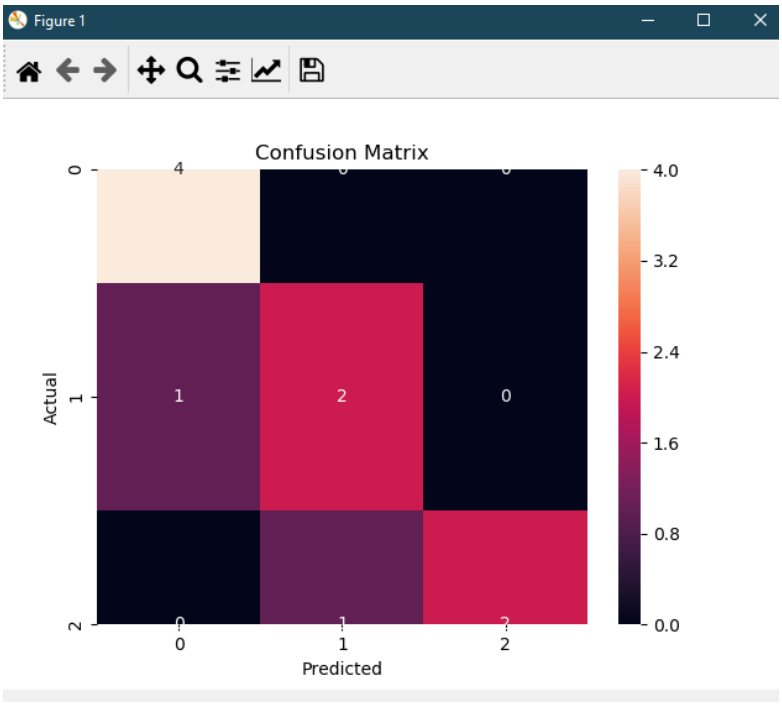
Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

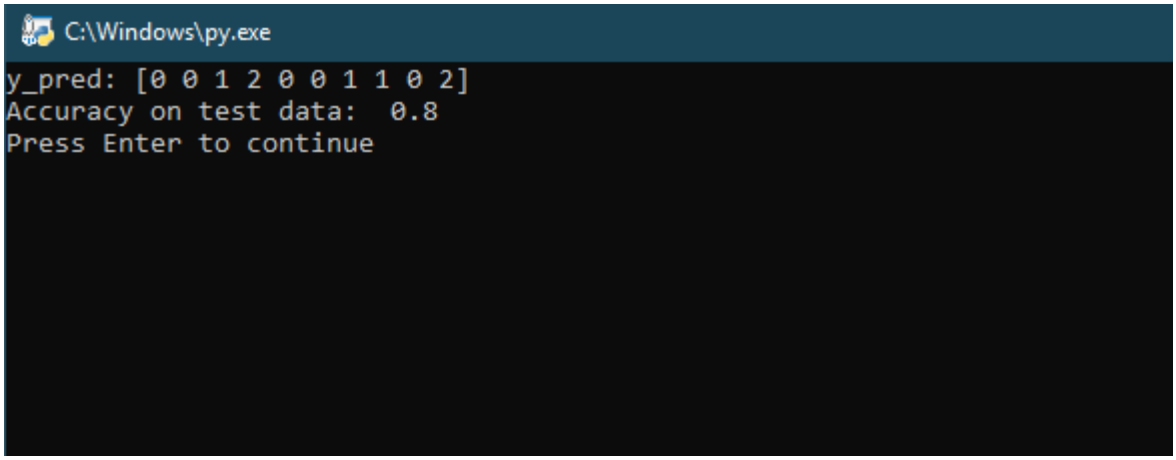
Seaborn:

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. Different plots in seaborn are heatmap, pairplot, countplot, distplot etc.

Confusion Matrix:



Accuracy Score:



Data splitting:

Splitting the data into train and test set which is 75%:25% ratio.

```
from sklearn.model_selection import train_test_split
X_train, X_test,  y_train, y_test = train_test_split(X, y, test_size=0.25, random_state = 0)
```

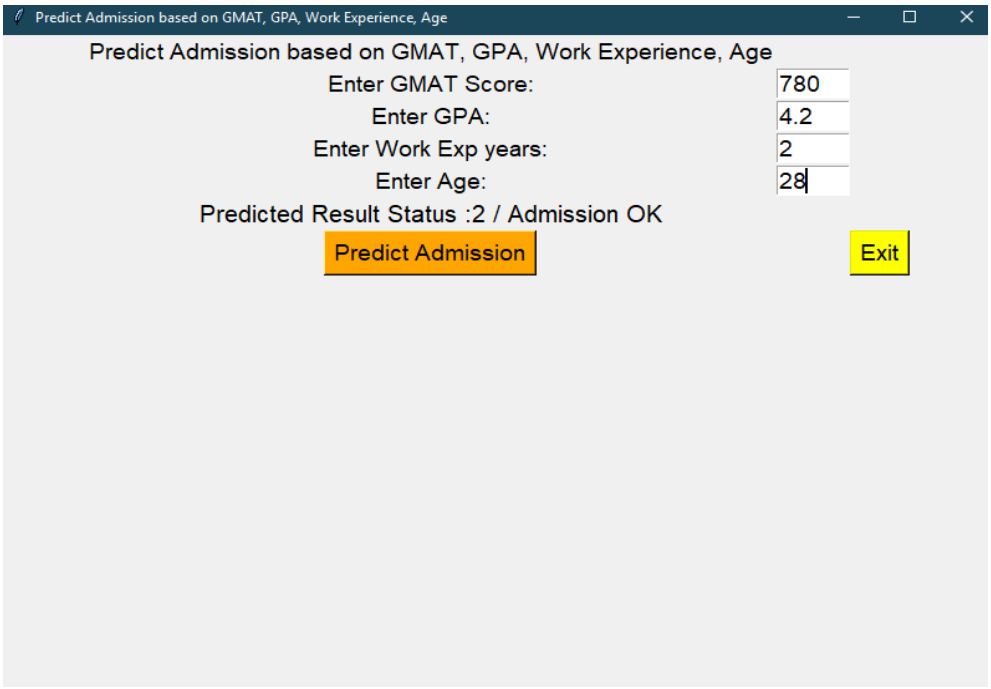
Model evaluation:

- Model is evaluated with the help of **test data set**.
- Test data set is independent of the train data set but follows same probability distribution as the training data set.
- Test data set is used to provide an unbiased evaluation of the final model fit on the training data set.

```python
# Program name : display_train_test_data.py

import pandas as pd

from sklearn.ensemble import RandomForestClassifier
df = pd.read_excel("candidates.xlsx")

X = df[['gmat', 'gpa', 'work_experience', 'age']]      # features 2D
y = df['admitted']                                     # label or targets 1D
from sklearn.model_selection import train_test_split
X_train, X_test,  y_train, y_test = train_test_split(X, y, test_size=0.25, random_state = 0)
clf = RandomForestClassifier(n_estimators = 100)
clf.fit(X_train, y_train)

print("X_train head (features):\n", X_train.head())
print("X_test head (features):\n", X_test.head())
print("y_train head (labels/targets):\n", y_train.head())
print("y_test head (labels/targets):\n", y_test.head())
x = input('Press Enter to Continue')

```

```
C:\Windows\py.exe
X_train head (features):
     gmat  gpa  work_experience  age
27   580  2.3               2   26
35   650  2.3               1   22
37   580  3.3               1   24
2    690  3.3               3   24
39   790  3.7               5   31
X_test head (features):
     gmat  gpa  work_experience  age
22   550  2.3               4   28
20   620  3.3               2   25
25   670  3.3               6   29
4    780  3.9               4   26
10   610  2.7               3   25
y_train head (labels/targets):
 27    0
35    0
37    0
2     1
39    2
Name: admitted, dtype: int64
y_test head (labels/targets):
 22    0
20    1
25    1
4     2
10    0
Name: admitted, dtype: int64
Press Enter to Continue
```

Make prediction: (Using Random Forest Regression):



Discussion

"Admission Prediction" is of regression problem in supervised learning because output variable(Salary) is a continuous value.

Aim: Our aim was to predict the chances of admission of a candidate according to his/her GMAT score,GPA score, age and work experience.

We could have used Logistic Regression model.

But by Random Forest Regression we achieved an accuracy of 80%.

Conclusion

Machine learning is an application of artificial intelligence (AI) that provides system the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

Our aim was to predict the chances of admission of a candidate according to his/her GMAT score, GPA score, age and work experience.

The prediction problem is of Regression problem in Supervised Learning.

Supervises learning is where we have input variables(X) and output variables(Y) and we use an algorithm to learn the mapping function from input to the output.

A Regression problem is when the output variable is a real or continuous value (here, it is 'no. of admissions').

Here, we can predict the admission of an candidate accurately depending upon their profile (GMAT score, GPA score, age, work experience).

References

- https://seaborn.pydata.org/
- https://www.geeksforgeeks.org/python-introduction-matplotlib/
- https://www.geeksforgeeks.org/introduction-machine-learning-using-python/