DEPARTMENT OF COMPUTER & INFORMATION SYSTEMS ENGINEERING
BACHELORS IN COMPUTER SYSTEMS ENGINEERING
Course Code: CS-324
Course Title: Machine Learning
Complex Engineering Problem
TE Batch 2019, Spring Semester 2022

Grading Rubric

TERM PROJECT

Group Members:

Student No.	Name	Roll No.
S1	**Muneeza Baig**	**CS-19015**
S2	**Zuhaira Abdullah Siddiqui**	**CS-19019**
S3	**Syeda Ghazia Hashmi**	**CS-19061**

CRITERIA AND SCALES				**Marks Obtained**		
				S1	**S2**	**S3**
Criterion 1: Does the application meet the desired specifications and produce the desired outputs? (CPA-1, CPA-2, CPA-3) [8 marks]						
1	2	3	4			
The application does not meet the desired specifications and is producing incorrect outputs.	The application partially meets the desired specifications and is producing incorrect or partially correct outputs.	The application meets the desired specifications but is producing incorrect or partially correct outputs.	The application meets all the desired specifications and is producing correct outputs.			
Criterion 2: How well is the code organization? **[2 marks]**						
1	2	3	4			
The code is poorly organized and very difficult to read.	The code is readable only to someone who knows what it is supposed to be doing.	Some part of the code is well organized, while some part is difficult to follow.	The code is well organized and very easy to follow.			
Criterion 3: Does the report adhere to the given format and requirements? **[6 marks]**						
1	2	3	4			
The report does not contain the required information and is formatted poorly.	The report contains the required information only partially but is formatted well.	The report contains all the required information but is formatted poorly.	The report contains all the required information and completely adheres to the given format.			
Criterion 4: How does the student performed individually and as a team member? (CPA-1, CPA-2, CPA-3) **[4 marks]**						
1	2	3	4			
The student did not work on the assigned task.	**The student worked on the assigned task, and accomplished goals partially.**	**The student worked on the assigned task, and accomplished goals satisfactorily.**	**The student worked on the assigned task, and accomplished goals beyond expectations.**			

Final Score = (Criterial_1_score x 2) + (Criteria_2_score / 2) + (Criteria_3_score x (3/2)) + (Criteria_4_score)
= _____

ACKNOWLEDGEMENT

The Machine Learning project “CGPA Prediction Model” has been accomplished by the joint efforts of following group members:

- **Muneeza Baig CS-19015**
- **Zuhaira Abdullah Siddiqi CS-19019**
- **Ghazia Hashmi CS-19061**

Under the guidance of **Ma’am Maria Waqas** & **Ma’am Mehwish Raza**, submitted to Computer Systems Engineering Department.

Examiner sign:

ABSTRACT

The technological advancements have influenced the society so as to take a leap towards success. Every technological reform is a small step towards advancement and progress of mankind. Developments in information technologies have also been impacting upon Educational Institutions. The introduction of technology in educational institutions results in efficient practices. Likewise, CGPA Prediction Model, aims to provides effective layout to students for prediction. The CGPA Prediction Model holds an objective system for predicting the best estimated CGPA for students using their 1^{st} year only, 1^{st} & 2^{nd} year only or 1^{st}, 2^{nd} & 3^{rd} years grades, providing them an early estimation for their goals and direction on the basis of CGPA. CGPA Prediction Model takes the information of grades of a student and predicts its CGPA on the basis of the provided grades. To put it succinctly, CGPA Prediction Model provides the students an ease of predicting their CGPA to keep a rack of their performances. Moreover, for achieving the best results, this model has been trained separately on a number of effective Machine Learning algorithms so to arrive at the one with the best prediction results.

TABLE OF CONTENTS

Data Preprocessing steps

i- Treating Null Values

The given Dataset of Grades contains null values, which can affect the accuracy of the dataset (for both training & testing) in order to predict the CGPA of a student. Thus, Null values have to be treated. Null values can be treated in a number of ways :

- Removing the entire rows containing null values
- Filling the null fields with Mean value of the column
- Filling the null fields with Mode value of the column
- Filling the null fields with Median value of the column

On the provided Dataset, Null values have been filled or replaced with **Mode** values of the respective columns.

ii- Encoding Techniques

The given Dataset of Grades contain string values; hence we need to encode them in order to convert them into processable numeric values. Encoding techniques are listed below :

- Ordinal Encoding
- Dummy Variable Encoding
- One hot Encoding

On the provided Dataset, Ordinal encoding technique has been applied to convert the string feature values into numeric feature values. The reason behind the application of ordinal encoding is that grades have a natural ranking hence can be given numbers according to the ranking. Following is the number assigned to each grade which is according to the GPA value. :

A +	4.0
A	4.0
A -	3.7
B +	3.4
B	3.0
B-	2.7
C +	2.4
C	2.0
C-	1.7
D +	1.4

D	1.0
F	0.0
WU	0.0
W	0.0

iii- Skewness of Dataset

The hypothesis behind the evaluation of skewness of the Dataset is :

Most of the parametric machine learning models like Linear Regression which has been used in one of the models implemented works well with the normally distributed data else the model fails to give accurate predictions.

Distplot() is used to visualize the parametric distribution of a dataset. Depending on the skewness of a density curve, we can quickly know whether the mean or median is larger in a given distribution. In particular:

- If a density curve is left skewed, then the mean is less than the median.
- If a density curve is right skewed, then the mean is greater than the median.
- If a density curve has no skew, then the mean is equal to the median.

There are a number of ways for treating skewness of Data. There are many negative values present in dataset. The technique applicable for treating negative values is **square root Transformation**. It is a transformation with a moderate effect on distribution shape.

With the application of square root transformation, we saw the reduction of skewness in the Dataset, but since the output of GPA prediction Model is supposed to be in the range of 0.06 CGPA (all values low) to 3.87 CGPA, the square root transformation could not hold the CGPA in the above-mentioned range. Thus, the square root transformation has not been applied to this CGPA model.

iv- P-value Significance

The hypothesis behind the evaluation of P-Value of the Dataset is :

P-Value deduce the significance of the Feature. If the P-Value is less than 0.05, only then the feature is significant. Since every feature (subject) is contributing to the CGPA and also significantly important for formulating the CGPA, due to this specific scope of the GPA prediction Model, the insignificant models have not been dropped out of the dataset.

Models and Machine Algorithms for Training

Model 1: predict final CGPA based on GPs of first year only.

Model 2: predict final CGPA based on GPs of first two years.

Model 3: predict final CGPA based on GPs of first three years.

1- Linear Regression.

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable

2- Regression Tree.

A regression tree is basically a decision tree that is used for the task of regression instead of classification and can be used to predict continuous-valued outputs instead of discrete outputs.

3- ANN

The purpose of using Artificial Neural Networks for Regression over Linear Regression is that the linear regression can only learn the linear relationship between the features and target and therefore cannot learn the complex non-linear relationship.

Artificial Neural Networks have the ability to learn the complex relationship between the features and target due to the presence of activation function in each layer.

4- Polynomial Regression

If we have non-linear data, then Linear regression will not be capable to draw a best-fit line and it fails in such conditions. Hence, we introduce polynomial regression to overcome this problem, which helps identify the curvilinear relationship between independent and dependent variables.

So, polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modeled as an nth degree polynomial in x.

Distinguishing Features

- A separate dashboard for the application is deployed using Voila which details and renders the Jupyter file for the application as a dashboard for better visualization. Voila is a library that is

used to directly convert your Jupyter notebooks (.ipynb files) into stand-alone interactive web-based dashboard applications.

- The main interface for the CGPA prediction system is created using Streamlit thus the prediction system is also rendered into an independent web application..

Tabular Comparison of the Models

The following table contains the model wise comparison of the models

Model 1	Model 2	Model 3
Model 1 consists of only the first-year courses. The data hence has only around **11** features to use for CGPA prediction.	Model 2 consists of courses from the first 2 years. The data hence has **22** features to use for CGPA prediction.	Model 3 consists of courses from all 3 years. The data hence has **33** features to use for CGPA prediction.
As is apparent from our analysis, that Model one, though predicts well still, lags behind in comparison to model 2 and 3	Model 2 performs considerably well as compared to model 1 but still not as well as model 3.	Model 3 outperforms models 1 and 2 due to the sheer amount of data the model utilizes for its prediction.
The best training and testing accuracy for Model 1 is **0.839** and **0.814** respectively.	The best training and testing accuracy for Model 2 is **0.908** and **0.906** respectively.	The best training and testing accuracy for Model 3 is **0.938** and **0.936** respectively.

The following table contains the algorithm wise comparisons of the model.

1- Linear Regression

	Model 1	Model 2	Model 3
Training accuracy	0.8397	0.9088	0.9380
Testing accuracy	0.81413	0.9066	0.9321

2- Regression Tree

	Model 1	Model 2	Model 3

Training accuracy	0.85550	0.9066	0.9256
Testing accuracy	0.64935	0.7526	0.7689

3- ANN

	Model 1	**Model 2**	**Model 3**
Training accuracy	0.8688	0.924	0.9597
Testing accuracy	0.7741	0.8234	0.9281

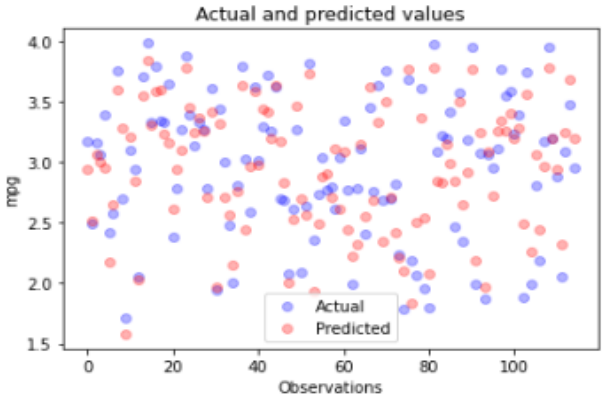
4- Polynomial Regression

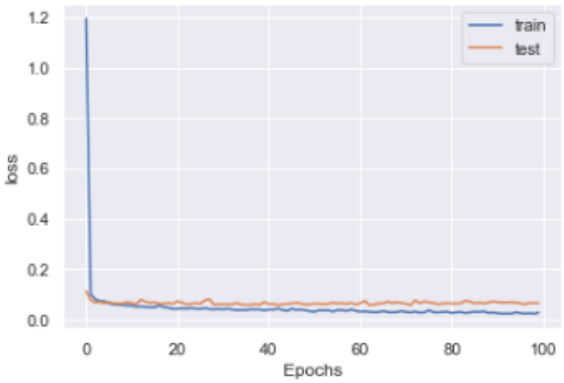
	Model 1	**Model 2**	**Model 3**
Training accuracy	0.9083	0.9856	1.0
Testing accuracy	0.7920	0.6987	0.9163

Graphical comparison of the Models

1- Model 1 – Using the Linear Regression Algorithm and ANN

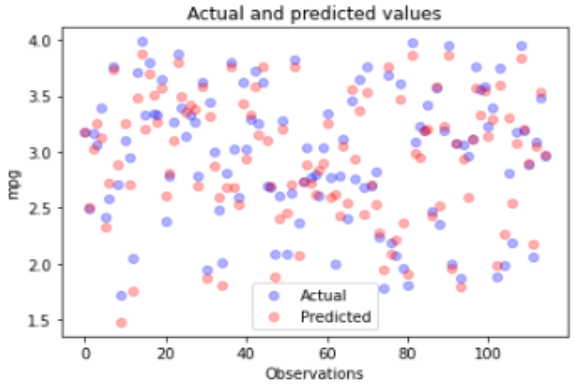
```
The training accuracy 0.8397211612644766
The testing accuracy 0.8141336830277861
```

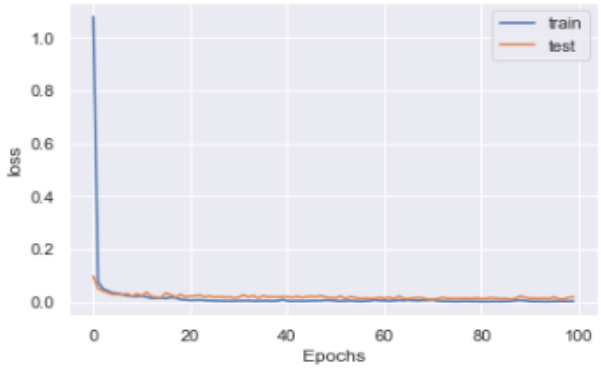




2- Model 2 – Using the Linear Regression Algorithm and ANN

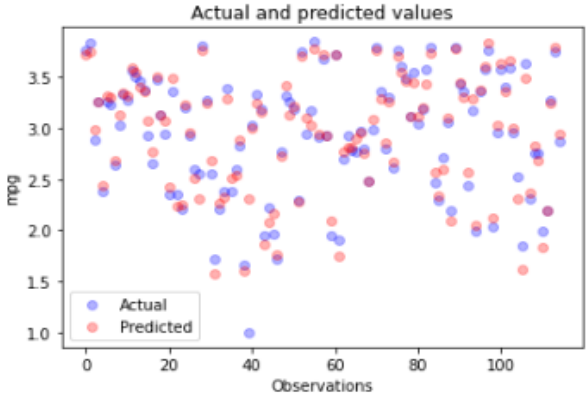
```
The training accuracy 0.9088336931042303
The testing accuracy 0.9066851498286942
```

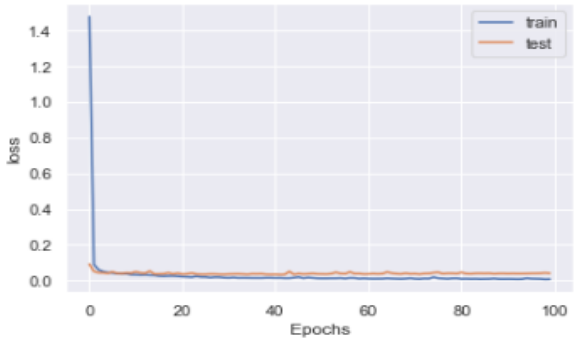




3- Model 3 – Using the Linear Regression Algorithm and ANN

```
The training accuracy 0.9384163315589276
The testing accuracy 0.9360703844054009
```





Performance of the implemented Machine Learning System

The implemented Machine learning models have been transformed to provide optimal predicted outputs. Issues such as overfitting and underfitting were observed but were tackled by properly splitting the data and handling distributions for training and testing using a random state. The dataset

was found to be slightly skewed were some individual features had higher P - values than the specific threshold of 0.05.

All these issues were catered to some degree depending on their nature and influence on the overall prediction system. All 3 models predict the final CGPA well, each predicting it more accurately than its former counterpart. Additionally, the models proved to be considered accurate and are in accordance with the general perception of accuracy standard, where all fall under the range of **81% - 93%.**