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	**Muhammad Zain Ammad**	**CS-19079**

CRITERIA AND SCALES				**Marks Obtained** **S1**
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)
= _____

DATA PREPROCESSING:

PREPROCESSING FOR ALL 3 MODELS:

Dropping Irrelevant Features:

Student ID has nothing to do with the prediction of CGPA so have dropped it from the dataset.

Dealing With Null Values:

I observed that 38 of the features of dataset contain null values. But 36 of the features (courses) contain very few null values (i.e. less than 15) so dropping 15-20 records from a dataset having 571 records doesn't harm us so I dropped the rows of courses which contain null values less than 15.

Encoding the categorical values:

For the encoding of categorical values, I just mapped the grades with relative grade points.

Remarks:

- Null values are removed from the dataset
    - CS-406 and CS-412 still contain null values because:
        - They have greater number of null values
        - We will not use 4th year courses to train our model
    - As a result, categorical value I (Incomplete) has also been removed from the dataset.
- Categorical values are encoded to numerical values
    - Some 3rd year courses still contain 2 categorical values i.e. W and WU which will be deal later
        - Why I haven't performed any preprocess for W and WU yet?
            - 1st and 2nd year courses are cleaned so I can train my 2 models without removing further records for W and WU

Result:

- We have cleaned our dataset to train our model 1 (prediction on basis of 1st year only) and model 2 (prediction on basis of 1st and 2nd year only) on 557 records
- Some more preprocessing required in order to train our model 3 (prediction on basis of 1st and 2nd and 3rd year)

PREPROCESSING FOR MODEL 3:

Dealing With W and WU values:

We have simply dropped the records which contain W and WU as they were very few in count (6 in total)

Result:

- Dataset has been cleaned to train our 3rd model on 546 data samples

SPLITTING THE DATASET:

In order to train 3 models, we need 3 sub datasets. What I have done is:

- Store CGPA in y (target variable) and removed the CGPA column from dataset so I only left with features.
- Then I split my dataset into 4 sub datasets
    - df_1: contain 1st year courses
    - df_2: contain 2nd year courses
    - df_3: contain 3rd year courses
    - df_4: contain 4th year courses (will not require that)
- To train model 1:
    - Simply use df_1 as m1_df (m1 data-frame)
- To train model 2:
    - Concatenate df_1 and df_2 as m2_df (m2 data-frame)
- To train model 3:
    - Concatenate df_1 df_2 and df_3 as m3_df (m3 data-frame)

MODELS & ALGORITHMS IMPLEMENTED:

MODELS:

I have implemented all the three models:

- Model 1: Prediction of CGPA on basis of 1st year courses only.
- Model 2: Prediction of CGPA on basis of 1st and 2nd year courses.
- Model 3: Prediction of CGPA on basis of 1st, 2nd and 3rd year courses.

ALGORITHMS:

The original dataset and the processed dataset contain 491 and 470 unique values of CGPA respectively so our target is continuous hence we have to use **Regression algorithms**.

Multiple Linear Regression:

Whenever we talk about regression, linear regression is the first choice that came to our mind. As we can see from the linear trend in data, linear regression may be a good fit on the data.

KNN Regression:

KNN regression is always a very good choice as far as it doesn't cost you in terms of computation. As our dataset wasn't too large so I used it.

Decision Tree Regression:

Just used it to have some hands-on in it. It may cause overfitting as we grow the tree deeper.

Random Forest:

It is a general assumption that random forest gives the best accuracy among all regression model so using this algorithm help me to make comparison among my algorithms.

DISTINGUISHING FEATURE:

The feature which can distinguish my project from others is Exploratory Data Analysis (EDA) and Visualization. Before doing data preprocess I tried to figure out my dataset by doing some exploratory data analysis so I got to know beforehand that my dataset contains 16 unique values including A+, A, A-, B+, B, B-, C+, C, C-, D+, D, F, I, W, WU, Nan. And by doing data visualization I got to know the trend in my data and according to which I have chosen the models to train, instead of doing hit and trial.

COMPARISON OF MODELS:

Calculating R^2 score on the models gives following result

	Model 01	Model 02	Model 03
Linear Regression	Training: 87%	Training: 95%	Training: 99%
	Testing: 72%	Testing: 94%	Testing: 99%
KNN Regression	Training: 87%	Training: 91%	Training: 95%
	Testing: 71%	Testing: 91%	Testing: 94%
Decision Tree Regression	Training: 87%	Training: 81%	Training: 83%
	Testing: 51%	Testing: 72%	Testing: 70%
Random Forest	Training: 97%	Training: 98%	Training: 99%
	Testing: 69%	Testing: 90%	Testing: 94%

PERFORMANCE OF ALGORITHMS:

PERFORMANCE ON MODEL 01:

All our models are causing overfitting because they are working fine on train data but fails to generalize and giving poor performance on testing data. As our dataset for model 1 is small and the applied algorithms are powerful enough so our model becomes over complicated. In order to avoid overfitting, we may use cross validation technique.

PERFORMANCE ON MODEL 02 & MODEL 03:

On both the models our all algorithms are working well except for KNN regression which is causing overfitting. Besides KNN all other models are working excellently and off course multivariate linear regression stands tall among them all (as we can see some linear trend in data as well through visualization).