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	Bisma Imran	CS-19016
S2	Izma Aziz	CS-19020
S3	Ayesha Aamir	CS-19026

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)

= _____

CONTENTS

STEPS OF DATA PRE-PROCESSING:

CHARACTER ENCODING:

As the features in our grades dataset are categorical, they cannot be used directly in most of the machine learning algorithm because mostly all machine learning algorithm are mathematical models, so we have used encoding technique to convert categorical features values to numeric values since the algorithms we have used are also mathematical. In this technique, the following steps have been performed:

1. A dictionary is created with the name “gp” for grade points that contains “grades” as keys and “grade points” as values using the given table as a guide.
2. A “replace” function with “gp” passed as argument is called on the dataset and as a result categorical data is converted into numerical (float) data.

Letter grade	Grade point	Letter Grade	Grade Point	Letter grade	Grade point
A+	4.0	**B-**	2.7	**C-**	1.7
A	4.0	**B**	3.0	**D**	1.0
A-	3.7	**C**	2.0	**D+**	1.4
B+	3.4	**C+**	2.4	**F**	0.0
W	0.0	**WU**	0.0	**I**	0.0

DROPPING UNNECESSARY FEATURES:

Since all three models depend on the first year, the second year or the third year courses, we have decided to drop the features which provided information on the fourth year courses.

This has been done using the drop method of Pandas library by passing the unnecessary columns, which were filtered out using the ‘-4’ pattern, and setting the axis as 0 to drop them row wise.

DEALING WITH NULL VALUES:

In any dataset, null values are problematic. Hence, we have first analyzed them to see which features have exactly how many null values. Since, the dataset was not that big, instead of just dropping the null values, we decided to first analyze them by extracting out rows where null value for a particular feature was present. In doing so, we discovered that many of the features had common null containing rows. For example, the row **CS-97045** contained null values for CY-105 and HS-105/12. Removing this row could effectively remove null values in CY-105 and HS-

105/12 as well. By using some helper functions (**valuable_info** and **print_proportions**), we determined the amount of information that each null containing row was providing for each of the three models.

If the information was more than 50% we decided to keep the row by just filling it with 0. If not, we dropped that row.

DUPLICATE ROWS:

We also checked the dataset for any duplicate rows and found none.

Scaling is also one of the data pre-processing techniques but in the given case it is not used as the values of features are already in a fixed range of grade point from 0 to 4.

MODELS AND ALGORITHMS IMPLEMENTED:

Two algorithms have been implemented for each of the three models that we have implemented in our project.

Model 1:

In this model we have used GPs of only first year to predict final CGPA at the end of fourth year. These courses include ***PH-121, HS-101, CY-105, HS-105/12, MT-111, CS-105, CS-106, EE-119, ME-107, CS-107***.

From domain knowledge, we deduced that a high Grade Point results in a higher CGPA. The relationship between the features in the grades dataset appears to be linear and the target variable in this dataset was continuous as well so the two algorithms that are implemented are regression models i.e. “SVR” and “Linear Regression”.

Model 2:

In this model we have used GPs of first two years to predict final CGPA at the end of fourth year. These courses include ***PH-121, HS-101, CY-105, HS-105/12, MT-111, CS-105, CS-106, EE-119, ME-107, CS-107, HS-205/20, MT-222, EE-222, MT-224, CS-210, CS-211, CS-203, CS-214, EE-217, CS-212, CS-215.***

The same two algorithms “SVR Algorithm” and “Linear Regression Algorithm” are implemented for this model since the dataset used here is also a subset of the actual dataset i.e. grades and thus the relationship appears the same.

Model 3:

In this model we have used GPs of three years to predict final CGPA at the end of fourth year. These courses include ***PH-121, HS-101, CY-105, HS-105/12, MT-111, CS-105, CS-106, EE-119, ME-107, CS-107, HS-205/20, MT-222, EE-222, MT-224, CS-210, CS-211, CS-203, CS-214, EE-217, CS-212, CS-215, MT-331, EF-303, HS-304, CS-301, CS-302, TC-383, EL-332, CS-318, CS-306, CS-312, CS-317.***

The same two algorithms "SVR Algorithm" and "Linear Regression Algorithm" are implemented for this model since the dataset used here is also a subset of the actual dataset i.e. grades and thus the relationship appears the same.

LINEAR REGRESSION:

Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

Since, the target variable was continuous and the relationship between variables according to our domain knowledge seemed linear, our first instinct was to try Linear Regression. Our decision proved fruitful since the train and test scores for all the models observed were more than average.

Model 3 gave the best train test score for Linear Regression which might have happened due to the increase in features.

SUPPORT VECTOR REGRESSION:

In contrast to OLS, the objective function of SVR is to minimize the coefficients more specifically, the L2-norm of the coefficient vector, not the squared error. The error term is instead handled in the constraints, where we set the absolute error less than or equal to a specified margin, called the maximum error, ϵ (epsilon).

SVR therefore gives us the flexibility to define how much error is acceptable in the model and will find an appropriate line to fit the data.

In sklearn, the default epsilon is 0.1 for SVR. Since, it was already giving us a high accuracy and by tweaking the parameter only a minor change was observed, we decided to keep the value as default.

All in all, the train and test scores for all three models turned out to be very high by using SVR.

DISTINGUISHING FEATURES:

For data cleaning we have implemented a function that tells us the valuable information contained by each of the rows with null values. The implementation of the function is as follows:

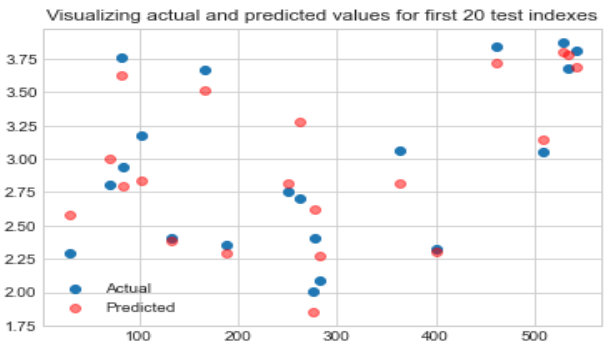
1. For each record with 1 or more null values, we have sum all the feature values that contain a not a null value.
2. Then we divide the value obtained in step 1 by the total number of courses in first, second and third year separately and cumulatively.
3. Then we divide the value obtained in step 1 by the total number of courses in first, second and third year cumulatively.
4. If the value obtained in step 2, by dividing it by FE courses, is much greater than the value obtained in step 3 then it indicates that the record is only useful for model 1 which contains only first year grade points.
5. So we drop that record from all other datasets for which it is not valuable.
6. And if the record is not valuable for any of the 3 years separately of cumulatively then we simply drop it from our datasets.

GRAPHICAL COMPARISON OF ALL THE MODELS:

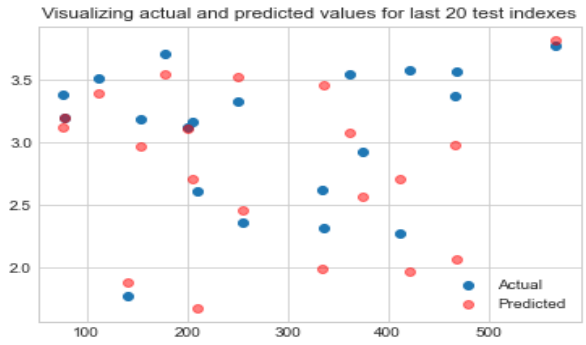
Following are the graphs of each model's actual and predicted values for first and last 20 indexes for each of the two models:

Model 1:

LINEAR REGRESSION:

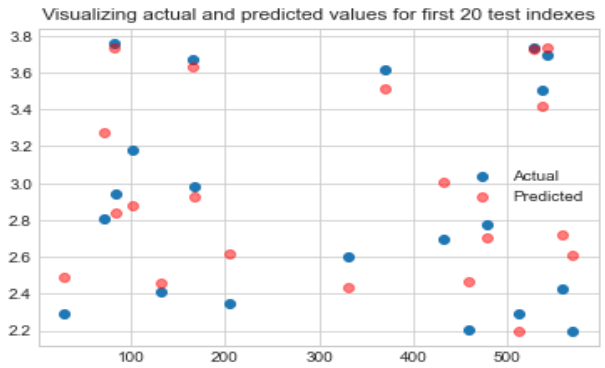


Linear regression for first 20 indexes

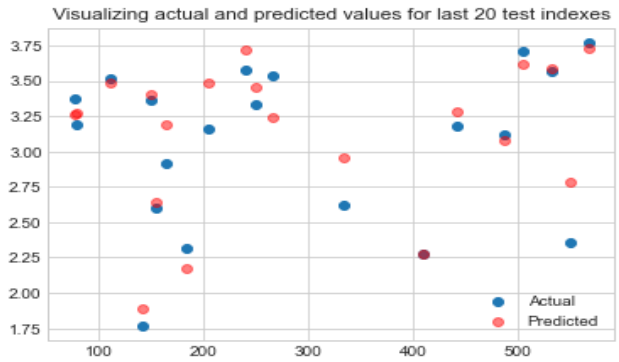


Linear regression for last 20 indexes

SVR:



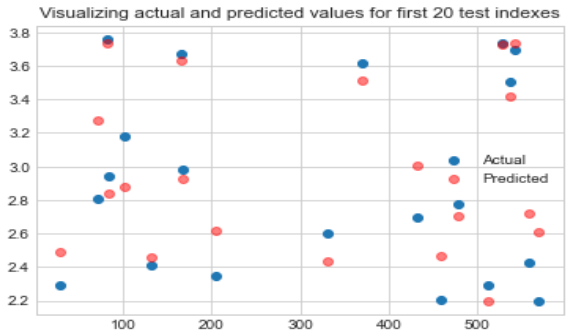
SVR algorithm for first 20 indexes



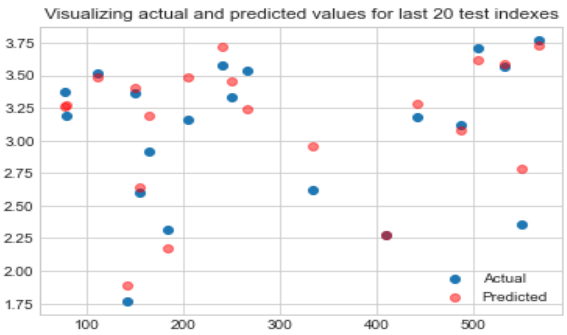
SVR algorithm for last 20 indexes

Model 2:

LINEAR REGRESSION:

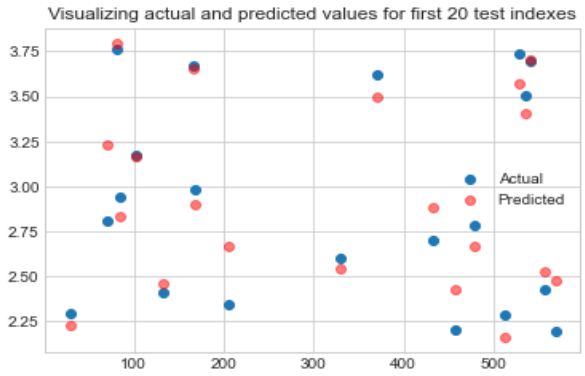


Linear regression for first 20 indexes

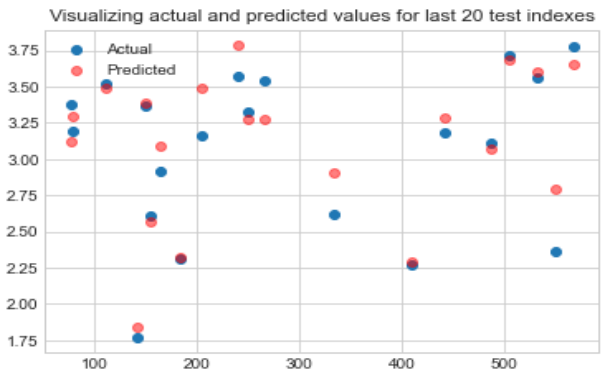


Linear regression for last 20 indexes

SVR:



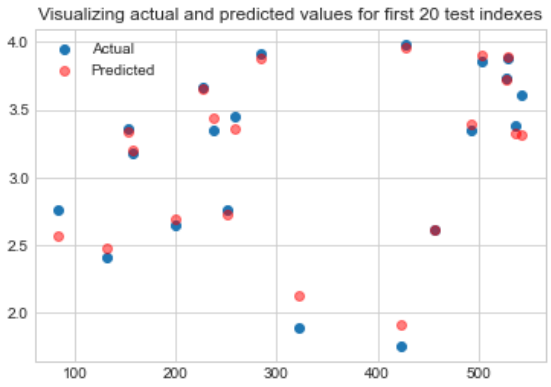
SVR algorithm for first 20 indexes



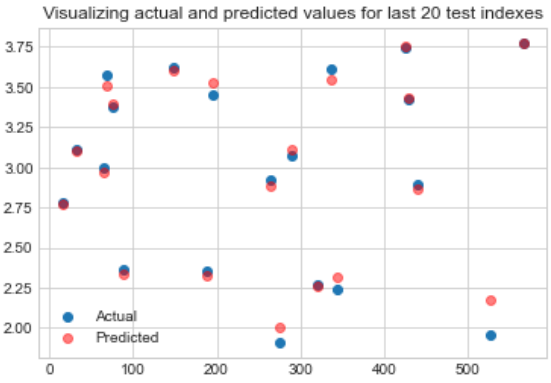
SVR algorithm for last 20 indexes

Model 3:

LINEAR REGRESSION:

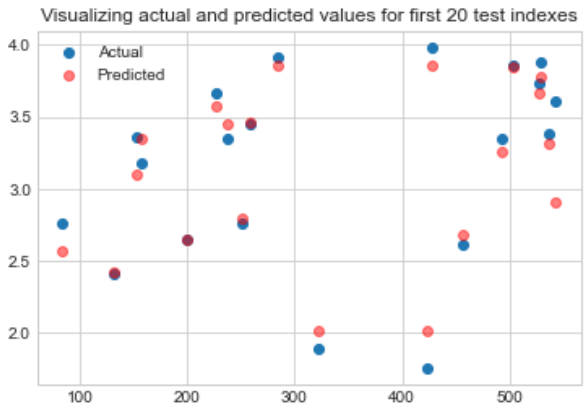


Linear regression for first 20 indexes

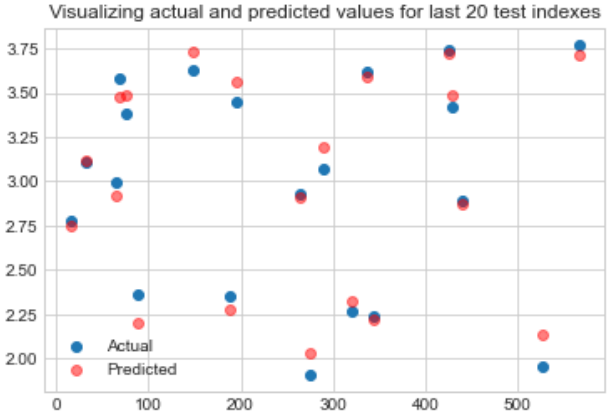


Linear regression for last 20 indexes

SVR:



SVR algorithm for first 20 indexes

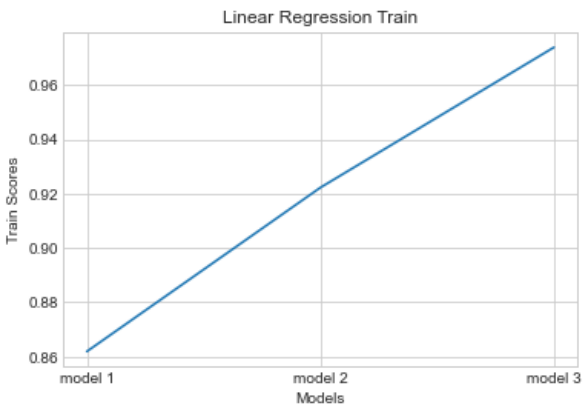


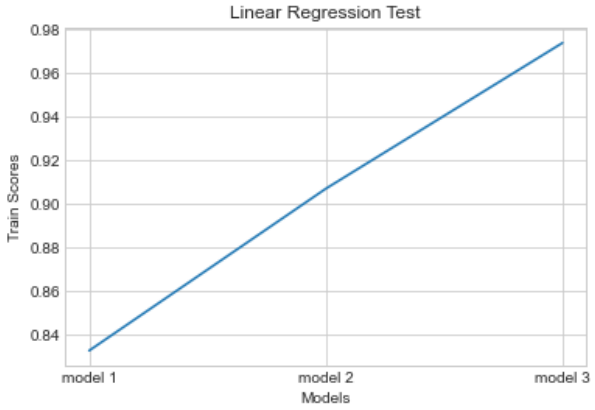
SVR algorithm for last 20 indexes

If we compare linear regression for all three models, then it is evident how actual and predicted values for model 3 are closer and even overlapping. Hence, model 3 gives better prediction as compared to model 2 and model 1 and similarly model 2 gives better prediction as compared to model 1. From this, we conclude that as we increase features in dataset, the performance of the algorithm also improves. Hence, model 3 is the best among all models.

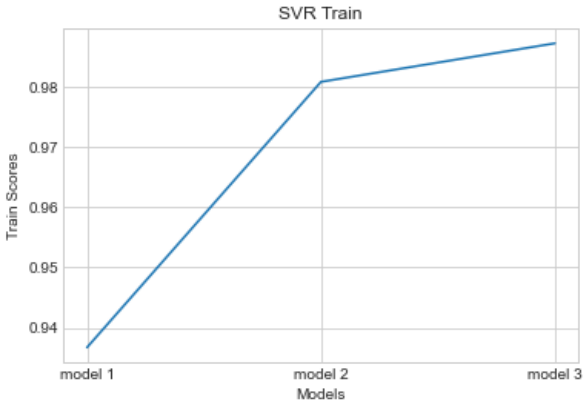
ACCURACY CHART FOR ALL THREE MODELS:

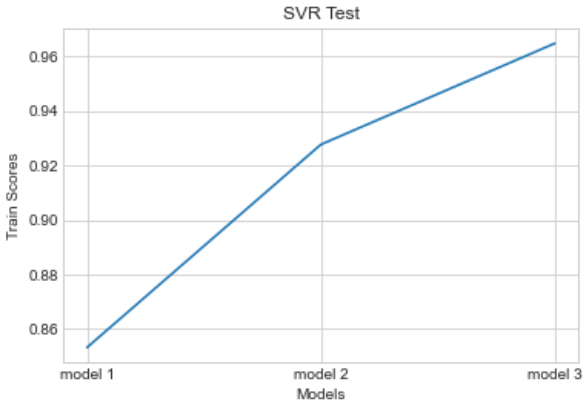
LINEAR REGRESSION:





SVR:





Scores improved by a greater factor in linear regression. From 86% in model 1 to almost 92% in model 2 to more than 96% in model 3. Whereas, in SVR, the scores improved by a small factor but were great overall. From 94% in model 1 to 98% in model 2 and more than 98% in model 3.

Comments on the performance of the implemented machine learning system, including issues like underfitting and overfitting, suggesting any techniques for improvement.

The problems like underfitting and overfitting are not found in our case as we have calculated accuracy of training and testing datasets for each model's both algorithm and found it to be over 80% for both training and testing. So, we can conclude that the overall performance of all three models is good and predictions can only be improved by using more data.