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	**Aiman**	**CS-19107**
S2	**Muhammad Umar**	**CS-19126**
S3	**Mir Shahnawaz**	**CS-19131**

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)

= _____

Prediction of Cumulative Grade Point Average

Data Preprocessing Steps

I. Dataset of CGPA is being provided (The_Grades_Dataset.csv).
II. VISUALISING THE DATA
III. Importing the required Libraries and their Machine Learning Algorithms.
- Pandas (Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data)
- NumPy (Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.
- Matplotlib.pyplot (It has a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: i.e creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.
- Scikit-learn (Simple and efficient tools for predictive data analysis)

IV. Specifying input features in variable x and target in varibale y.
FEATURE ENGINEERING
   a. Identify Anomalies/ Missing Data filling Empty spaces and Not a Number (NaN) in a cell using mode technique.
   b. Encodes the given data in an ordinal values from 1 - 15, where 1 represents least weightage and 15 represents highest weightage of a grade point, replaces all alphabetic grades with integer values.
   c. Establish a baseline model
   d. Splitting dataset in to two where 80% used for training and 20% for testing.

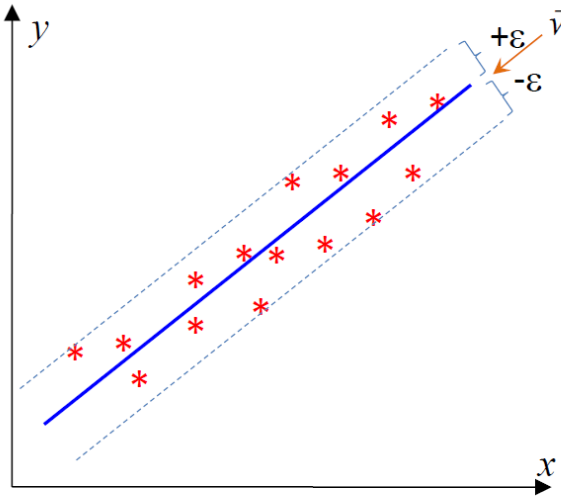
Machine Learning Algorithms

We have selected Model 2 and 3 for prediction of CGPA from the given models.

Following Algorithms is being implemented on a provided dataset.

- ### Support Vector Machine – Support Vector Regression

  Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.



Find $f(\vec{x})=\vec{w}\cdot\vec{x}+b$

by minimizing $\frac{1}{2}\|\vec{w}\|^2$ subject to constraints:

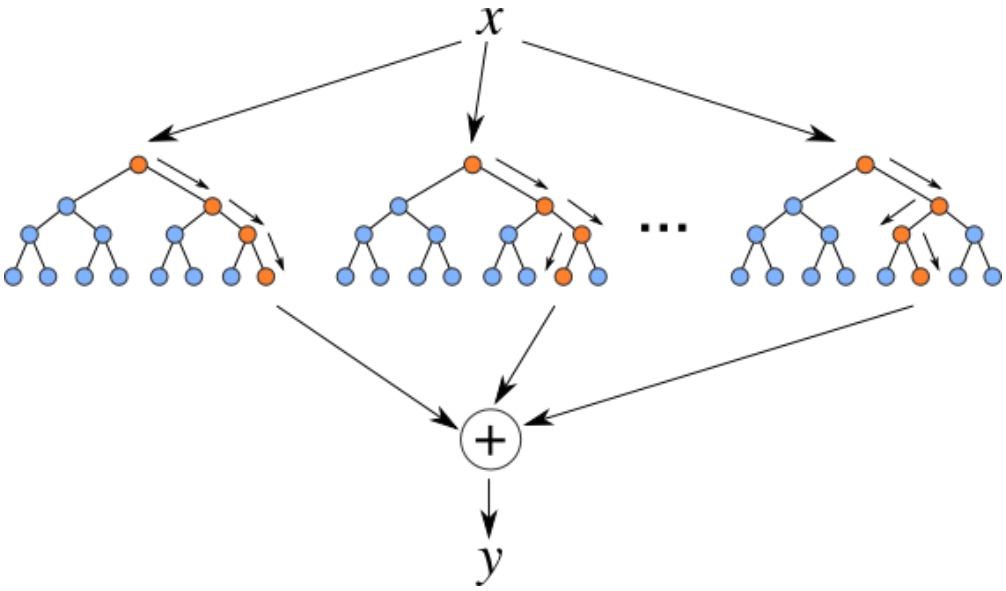
$$y_i-(\vec{w}\cdot\vec{x}+b)\le\varepsilon$$
$$y_i-(\vec{w}\cdot\vec{x}+b)\ge-\varepsilon$$

for $i=1,\ldots,N$.

I.e., difference between y_i and the fitted function should be smaller than ε and larger than -ε ⇔ all points y_i should be in the "ε-ribbon" around the fitted function.

- **Random Forest Regressor**

Random forest is an ensemble learning algorithm based on decision tree learners. The estimator fits multiple decision trees on randomly extracted subsets from the dataset and averages their prediction. It is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the `max_samples` parameter if `bootstrap=True` (default), otherwise the whole dataset is used to build each tree.



- **Linear Regression**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

$$y = \theta_1 + \theta_2 . x$$

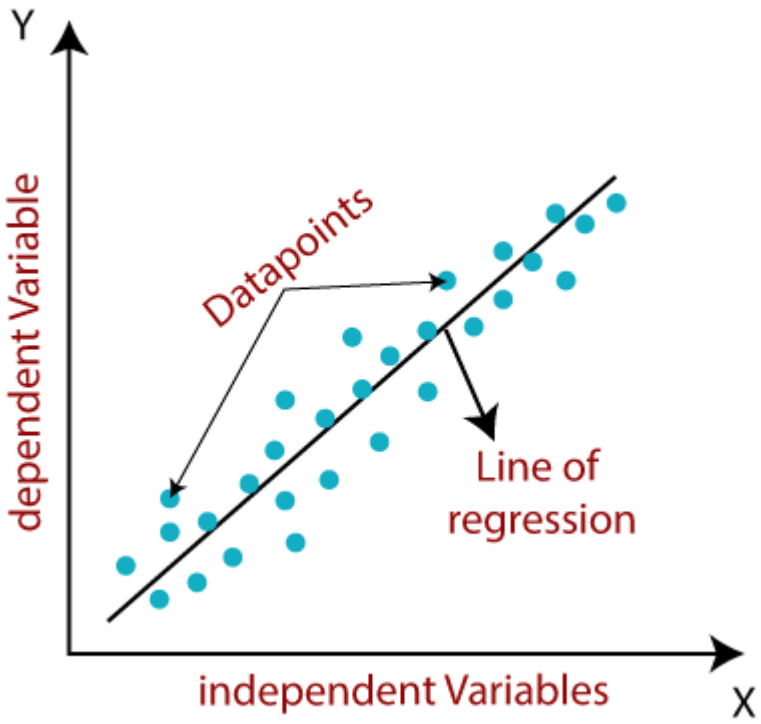
While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)
When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.
θ1: intercept
θ2: coefficient of x
Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

Cost Function (J):
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

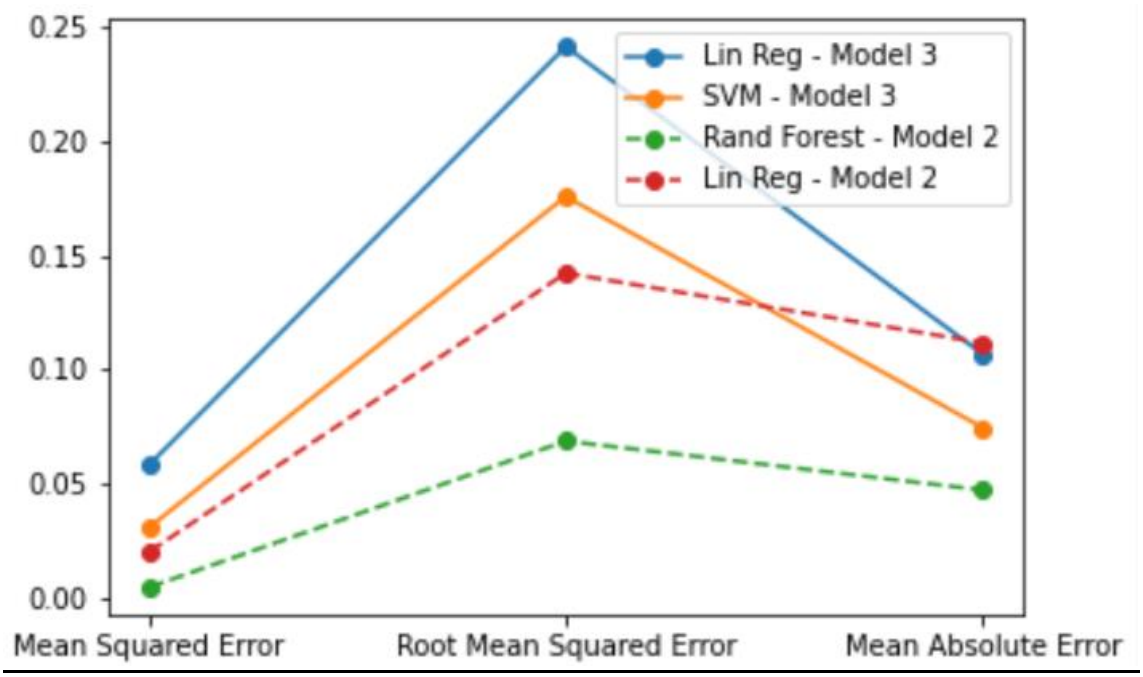
$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Cost function (J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).



Comparison of Algorithms



Score of Models

Algorithms	Scores	Model 3
Linear Regression	Training	95.2%
	Testing	87.5%
SVM-R	Training	92.7%
	Testing	92.4%

Algorithms	Scores	Model 2
Random Forest	Training	98.2%
	Testing	98.5%
Linear Regression	Training	89.7%
	Testing	94.9%

Comments

From the above graph we can conclude that the model 2 which has a data of first two year performed well as it has less MSE(Mean Squared Error) RMSE(Root Mean Squared Error) and MAE(Mean Absolute Error).
As input grades given same to all models and Model 2 performance is good as compare to Model 3.