Student marks prediction

ML project CO-327

By:- Yash Saini
Team:- 45
Roll no:- 2K18/EC/195
Submitted to:- Mrs Juhi Jain

Motivation

To reduce burden from those students who couldn’t able give exams

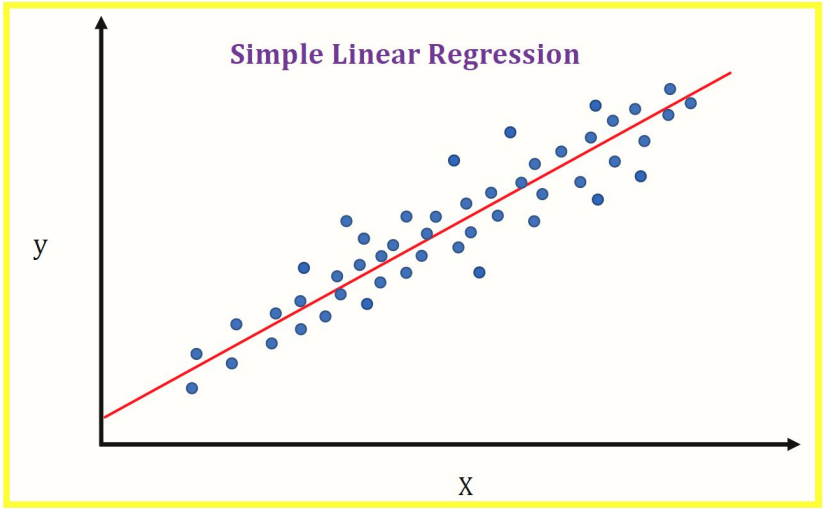
Students should be given a choice to choose whether to go with a retest or go with the marks predicted by the model itself.

1st	8.8
2nd	9
3rd	**6**
4th	8.4

Objective

To predict CGPA using Linear Regression model.

Training the model with the previous cg of students and making prediction and



The actual model will look like the one above where x axis will be 3rd sem cgpa and y will be 4th sem cgpa , which we will predict.

y_train = x_train * m + c

Where **m:- slope** and **c:- y intercept**

4.1 :- Dataset

Dataset comprised of around 1500 Students With their roll number and 3rd and 4th semester cgpa.

Branch:- AE, BT, CO, EC, EN, EP, IT ,MC, ME, PE, PS, SE

	roll_no	cg3	rollno	cg4
0	2K18/AE/003	4.64	2K18/AE/003	7.45
1	2K18/AE/004	4.55	2K18/AE/004	5.82
2	2K18/AE/005	7.64	2K18/AE/005	7.36
3	2K18/AE/006	9.27	2K18/AE/006	9.82
4	2K18/AE/007	6.91	2K18/AE/007	7.18
...	...	...	...	...
1541	2K18/SE/133	6.55	2K18/SE/133	6.82
1542	2K18/SE/135	7.09	2K18/SE/135	7.45
1543	2K18/SE/136	7.55	2K18/SE/136	9.09
1544	2K18/SE/137	7.91	2K18/SE/137	8.91
1545	2K18/SE/138	1.64	2K18/SE/138	2.50

1546 rows × 4 columns

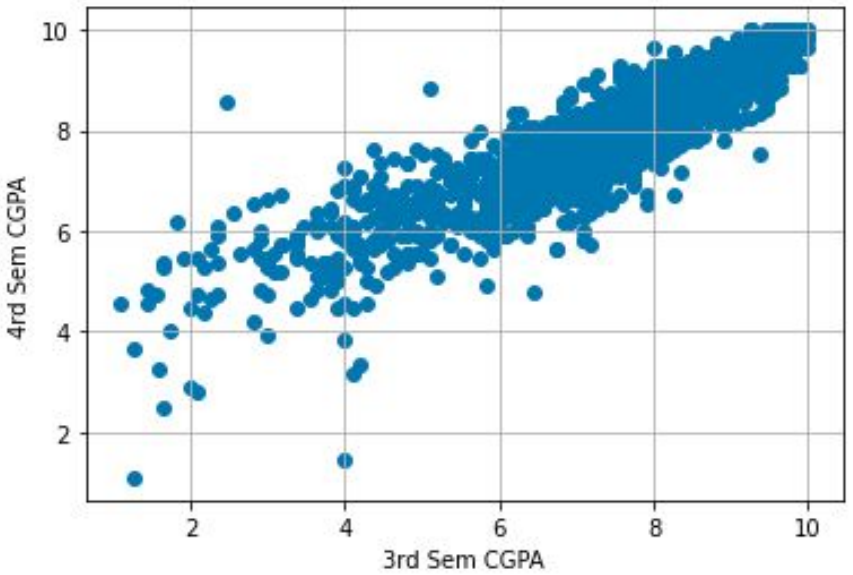
3rd Sem cgpa, also 1st and

2nd cgpa be added for prediction.

Dependent variable :- 4th sem cgpa.

Distribution of all the branches are same

Therefore, entire dataset is used as one.



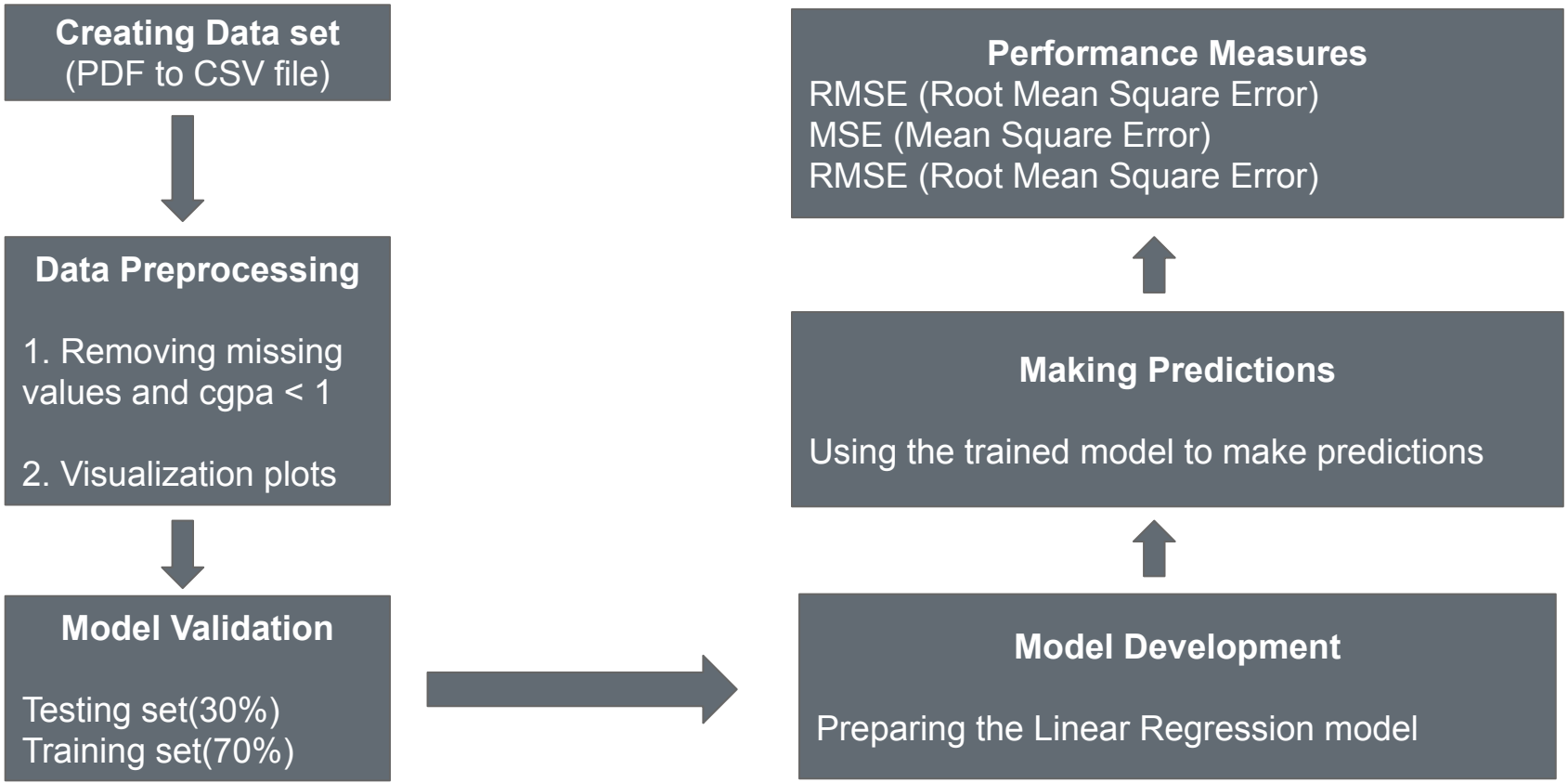
4.4 Preprocessing and Data analysis techniques

Handling missing values and removed instances with less than 1 cgpa.

Scatter plot to visualize dataset

Density plot to visualize the accuracy and performance of students.

5 Research Methodology

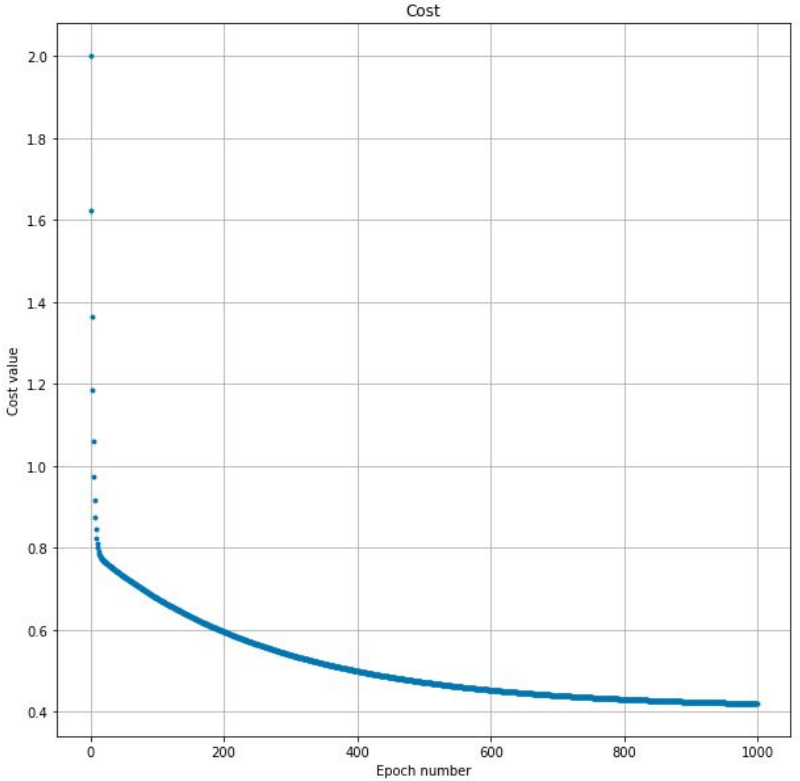


5.1 Data Preprocessing techniques

1) Handling missing values :- Removing the entire row if any of the cgpa is missing.
2) Removing rows with cg less than 1
3) Scatter plot to visualize distribution of grades based on different branch
4) Histogram plot to visualize difference of grades in both semesters

5.2 Data Analysis

1) Used Gradient Descent to minimize the cost function.



5.3 Validation Techniques

Split data into training and testing set in the ratio of 7:3

```
from sklearn.model_selection import train_test_split

X = df['cg3']
y = df['cg4']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

```
(1071,) (459,) (1071,) (459,)
```

5.4 Performance Measures

RMSE :- Root Mean Squared Error

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

MAE :- Mean Absolute error

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

R2 Score:- The proportion of the variance in the dependent variable that is predictable from the independent variable

Formula :- total Variance explained by model/ Total Variance

MBE :- Mean bias error, mean of difference between the actual value and predicted values

Formula :- (y_test - y_pred).mean()

Result

Mean Absolute percentage error	5.966871966496103 %
Mean Square error	0.33045853330288827
Root mean square percent error	9.080340517959428
R2 Score	0.7969448189729104
Mean Bias Error	- 0.006106677635694208

Thank You!