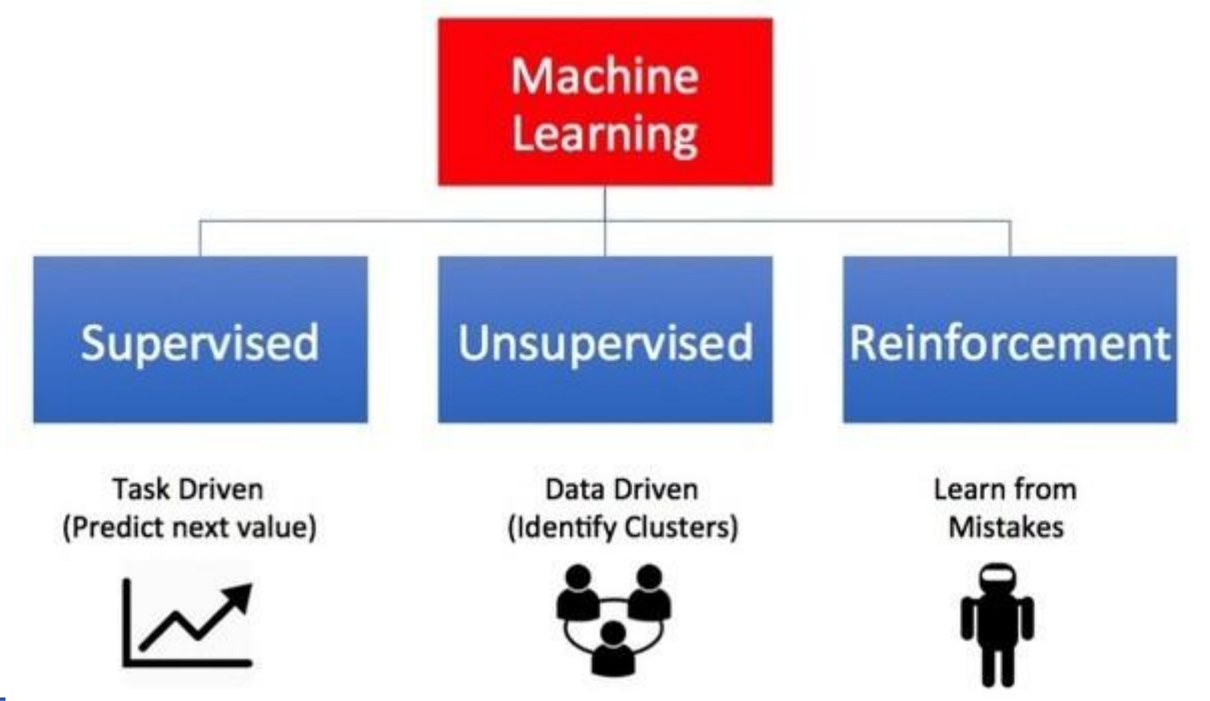
MACHINE LEARNING
Linear Regression

- Trainer : Sujata Mohite
- Email: sujata.mohite@sunbeaminfo.com

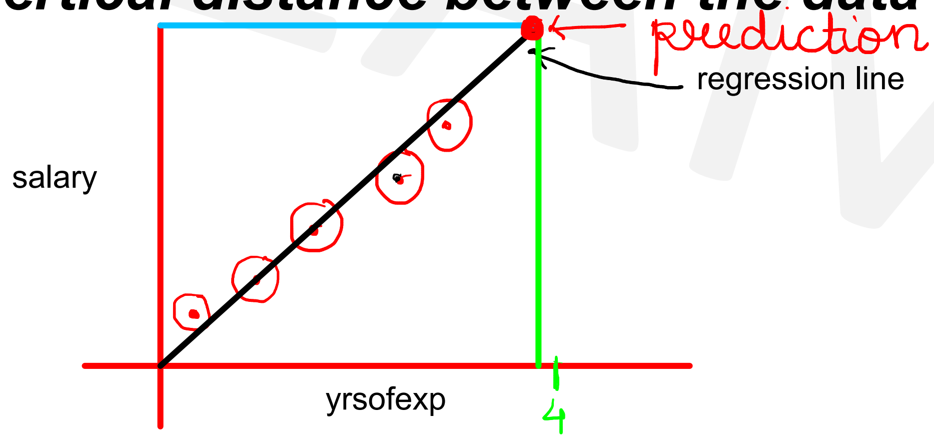
Supervised
Unsupervised
Reinforcement Learning



What is Regression analysis ?

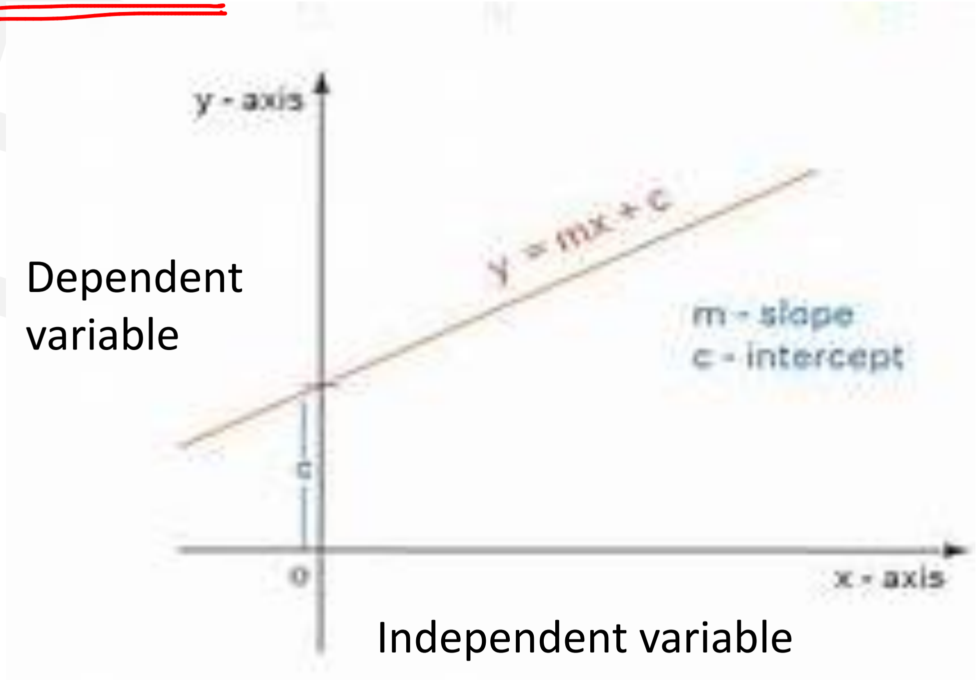
- Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.
- Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.
- It predicts continuous/real values such as **temperature, age, salary, price,** etc.
- In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data.
- ***Regression shows a line or curve that passes through all the data points on target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum.***

yrsofexp	salary
1	10k
1.5	20k
2	30k
2.5	40k
3	50k
3.5	60k
4	?



Linear Regression

- The data in Linear Regression is modelled using a straight line
- Linear regression is a statistical regression method which is used for predictive analysis.
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used with continuous variable
- It gives a future value as an output

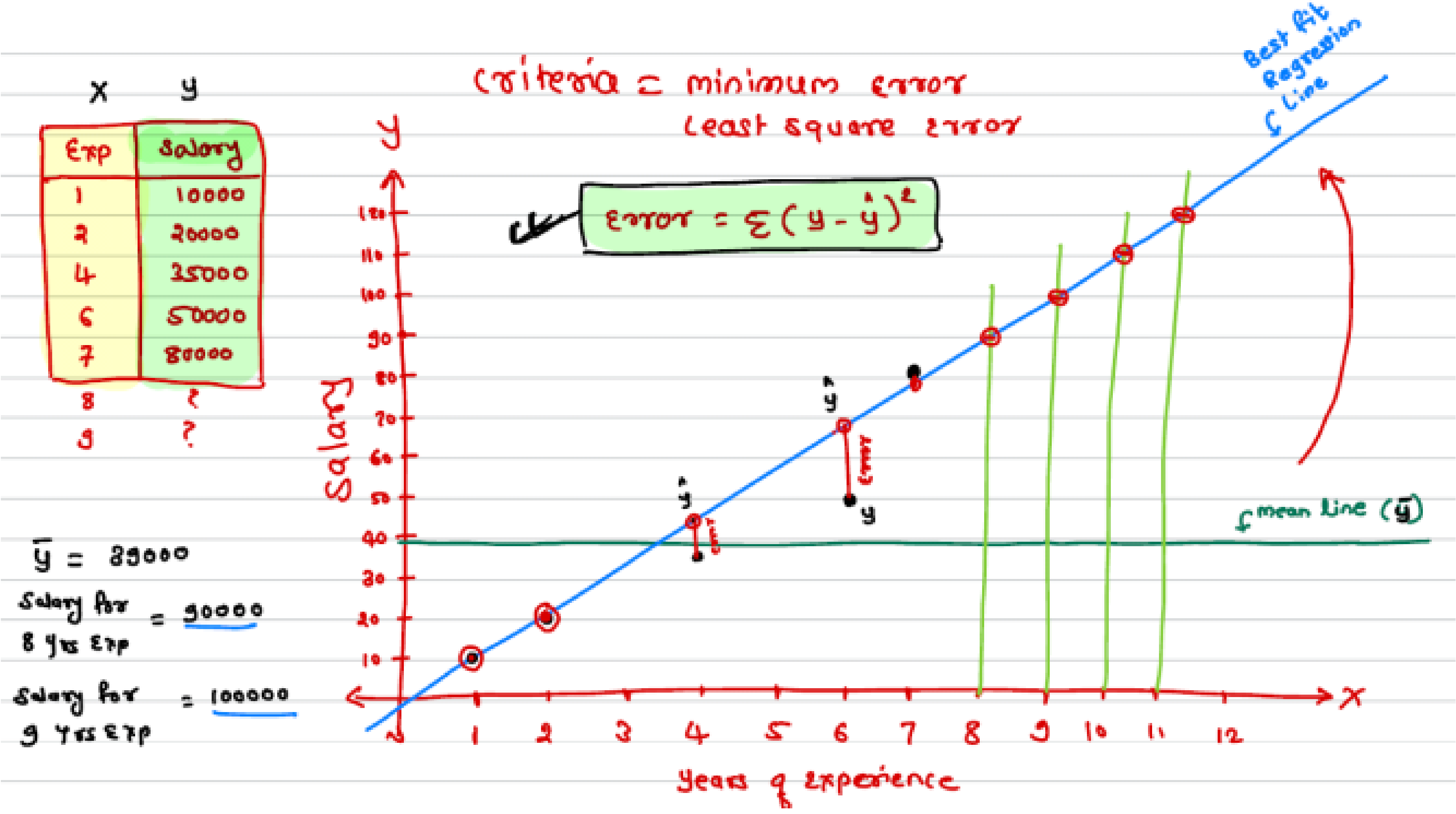


Dependent variable

Independent variable

Linear Regression

- Predicting the salary of an employee on the basis of the year of experience



x Exp	y Salary
1	10000
2	20000
4	35000
6	50000
7	80000
8	?
9	?

$\bar{y}$ = 39000

Salary for 8 yrs Exp = 90000

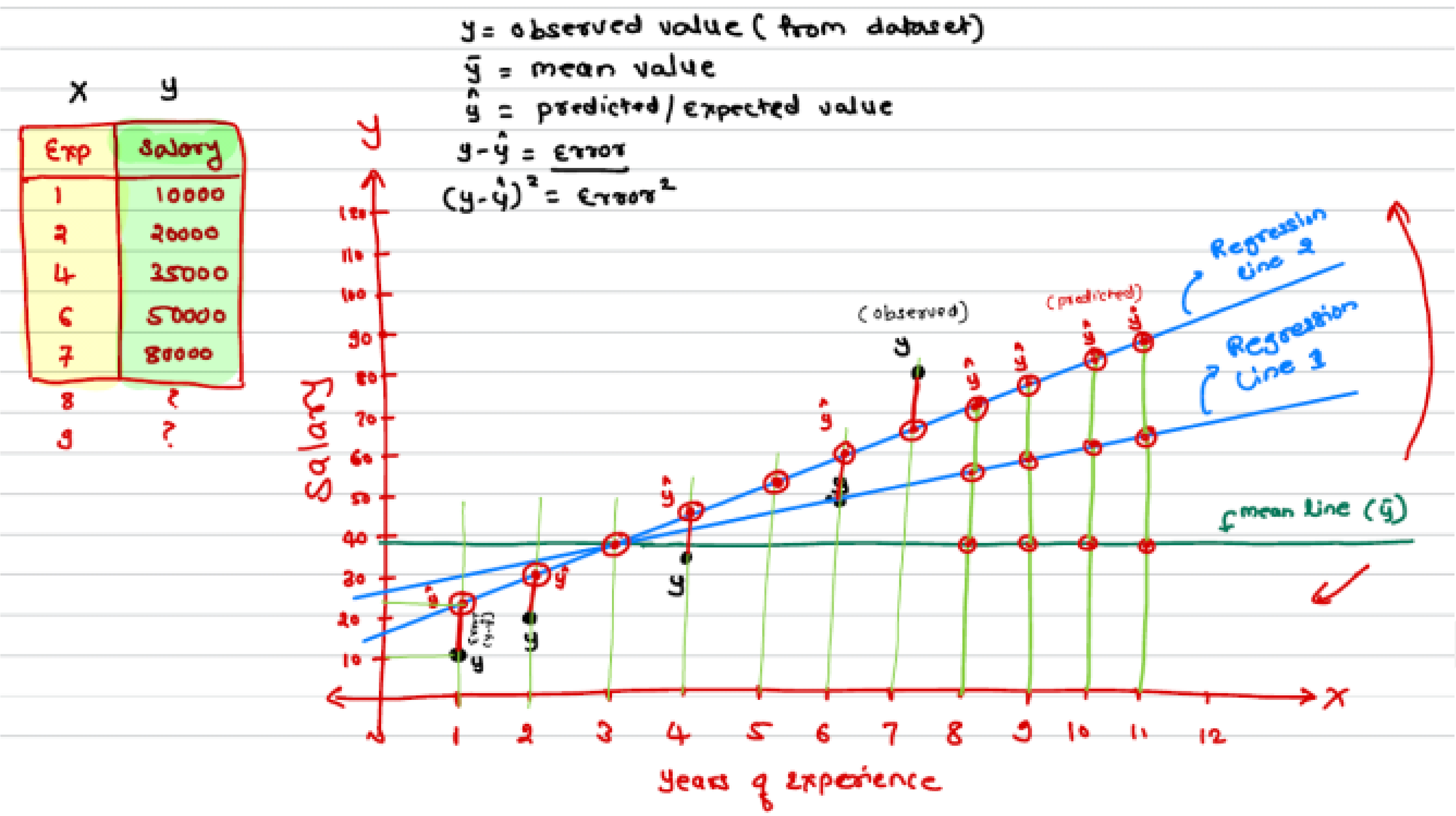
Salary for 9 yrs Exp = 100000

criteria = minimum error

least square error

$$\text{Error} = \Sigma(y - \hat{y})^2$$

Linear Regression



y = observed value (from dataset)

$\bar{y}$ = mean value

$\hat{y}$ = predicted / expected value

$y - \hat{y}$ = error

$(y-\hat{y})^2$ = error^2

x	y
Exp	Salary
1	10000
2	20000
4	35000
6	50000
7	80000
8	?
9	?

Least Square Method

- A form of mathematical regression analysis used to determine the line of best fit for a set of data, providing a visual demonstration of the relationship between the data points
- Each point of data represents the relationship between a known independent variable and an unknown dependent variable
- The least squares method provides the overall rationale for the placement of the line of best fit among the data points being studied
- It aims to create a straight line that minimizes the sum of the squares of the errors that are generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value, and the value anticipated, based on that model
- It begins with a set of data points to be plotted on an x-and y-axis graph
- An analyst using the least squares method will generate a line of best fit that explains the potential relationship between independent and dependent variables.

Regression Model Evaluation Metrics

- For evaluation of regression model, following metrics are used
- MAE – avg of absolute difference between actual & predicted value in a dataset

  - as it ignores signs, so we go for MSE
- MSE - avg of squared difference between the original and predicted values in a dataset.
- - lower MSE higher is accuracy better the model as it goes towards zero. As values are big due to square so we go for RMSE
- RMSE – Better to use as values are smaller and also does not ignore signs(directions +/-)
- - RMSE decreases model performance increases means accuracy increases.
- - RMSE does not have a benchmark(scale) to compare
- R2 - It's a way to find how good a model is. It shows difference between regression line and mean line.
- Adjusted R2 – difference between expected value and means value.
- In real world we don't use adjusted R2 score.

Mean Absolute Error (MAE)

- The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction (absolute)
- It measures accuracy for continuous variables
- The MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation $|y - \hat{y}|$
- The MAE is a linear score which means that all the individual differences are weighted equally in the average

$$\frac{\sum |y - \hat{y}|}{n}$$

Mean Squared Error (MSE)

- In statistics, the mean squared error (MSE) or mean squared deviation (MSD) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the error
- The average squared difference between the estimated values and the actual value
- MSE is a risk function, corresponding to the expected value of the squared error loss
- The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate
- The MSE is a measure of the quality of an estimator
- As it is derived from the square of Euclidean distance, it is always a positive value with the error decreasing as the error approaches zero

$$\frac{\sum (y - \hat{y})^2}{n}$$

Root Mean Squared Error (RMSE)

- RMSE is the most popular evaluation metric used in regression problems
- It follows an assumption that error are unbiased and follow a normal distribution
- Here are the key points to consider on RMSE:
- The power of ‘square root’ empowers this metric to show large number deviations
- The ‘squared’ nature of this metric helps to deliver more robust results which prevents cancelling the positive and negative error values
- It avoids the use of absolute error values which is highly undesirable in mathematical calculations
- When we have more samples, reconstructing the error distribution using RMSE is considered to be ore reliable
- RMSE is highly affected by outlier values. Hence, make sure you’ve removed outliers from your data set prior to using this metric.
- As compared to mean absolute error, RMSE gives higher weightage and punishes large errors

R-Squared (R2)

- We learned that when the RMSE decreases, the model’s performance will improve
- But these values alone are not intuitive
- When we talk about the RMSE metrics, we do not have a benchmark to compare
- This is where we can use R-Squared metric
- In other words how good our regression model as compared to a very simple model that just predicts the mean value of target from the train set as predictions

Performance Measures

Regression

- RMSE (Root Mean Squared Error) :square root of Mean Squared error
- MAE (Mean Absolute Error) : the average of the absolute difference between the actual and predicted values in the dataset
- MSE :the average of the squared difference between the original and predicted values in the data set

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

Where,

$\hat{y}$ – *predicted value of y*

$\bar{y}$ – *mean value of y*

The lower value of MAE, MSE, and RMSE implies higher accuracy of a regression model

Thank You!!