Assignment No: 1

Title: Linear Regression

Problem Statement: The following table shows the result of recently conducted study on the correlation of no. of hours spent driving with the risk of developing quite backache. Find the equation of best fit line for this data.

No. of hours spent	Risk Swe on Scale
No. of hours spent driving (X)	of 0-100 (y)
10	95
9	80
2	[0
15	50
10	45
6 ale Vanta de la	98
an established the second has	38
16	93

Objective: Students should be able to build a linear model for the given data.

Outcome: After completion of this assignment, students are able to understand how to find correlation between variables and how to calculate accuracy of linear model.

S/w requirements: Anaconda, python, sklearn, linun os N/w requirements: is processor, 4 gb sam Concept related theory: Uncar regression: Regression analysis is one of the nost midely used statisfical techniques. It estimates relationship among a dependent variable and independent variables. What is linear negression -In souse and effect relationship, the independent variable is the variable is the effect. Least square regression is the nethod of predicting the value of dependent variable based on value of independent variable. n. 2) The least square regression line: Given the random sample of observations population regression line is estimated by: j = bo + bin, where bo is constant be is regression coefficient, n is value of independent variable of is value of dependent variable.



To find the relationship between dependant and independant

The formula to find correlation using pearson correlation is -

8 = covariance (x, y)
84d. dev. (x) * std. dev (y)

3) How to define regression line?

We know regression model:

ŷ = bo + bix

The formula for calculating be & bo.

 $b_1 = \underbrace{\Xi(x_i - \overline{x}) + (y_i - \overline{y})}$

E (71 - 7)

bo = y - bix

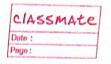
4) Accessing the model

For access the model, usually RSE (Residual Standard Error) and Restatistics are used.

$$RSE = \int \frac{1}{n-2} R88 = \int \frac{1}{n-2} \frac{2}{i=1} (yi - \overline{y}i)^2$$

$$R^2 = TSS - RSS = 1 - RSS$$
 $TSS = TSS$

- The first error natrin is simple to understand, the lower the residual error, the better the model fits the data.
- ii) As for the R2 nothing it measures the propostion of variability in the target that can be emplained using a future X. so, assuming the linear relationship if feature X can predict the target, then the propostion is high and R2 value will be close to I. If the opposite is true, then R2 value will be closer to D.
- 5) What is best line?
- i) The best line is a straight line that represents the
- ii) This line may pass through all-the points or none or att some of points.
- iii) Best line is deterrined from the matric such as



1. Calculate average of X variable. 2. Calculate différence between each X and Hlgorithm average X. 3. Square the differences & add it all up. SS,
4. Calculate average of Y variable.
5. Multiply difference of X & Y from
respective averages and add them.
This Sbry
6. Using 35xx and SSry, Calculate
; stercept by subtracting 55xy / 55xx t

Aug (n) from Aug (y). Linear regression equation: 4.58 x + 12.58 Conclusion: Thus, we learned to find the trend of data using X & Y variable using linear regression.