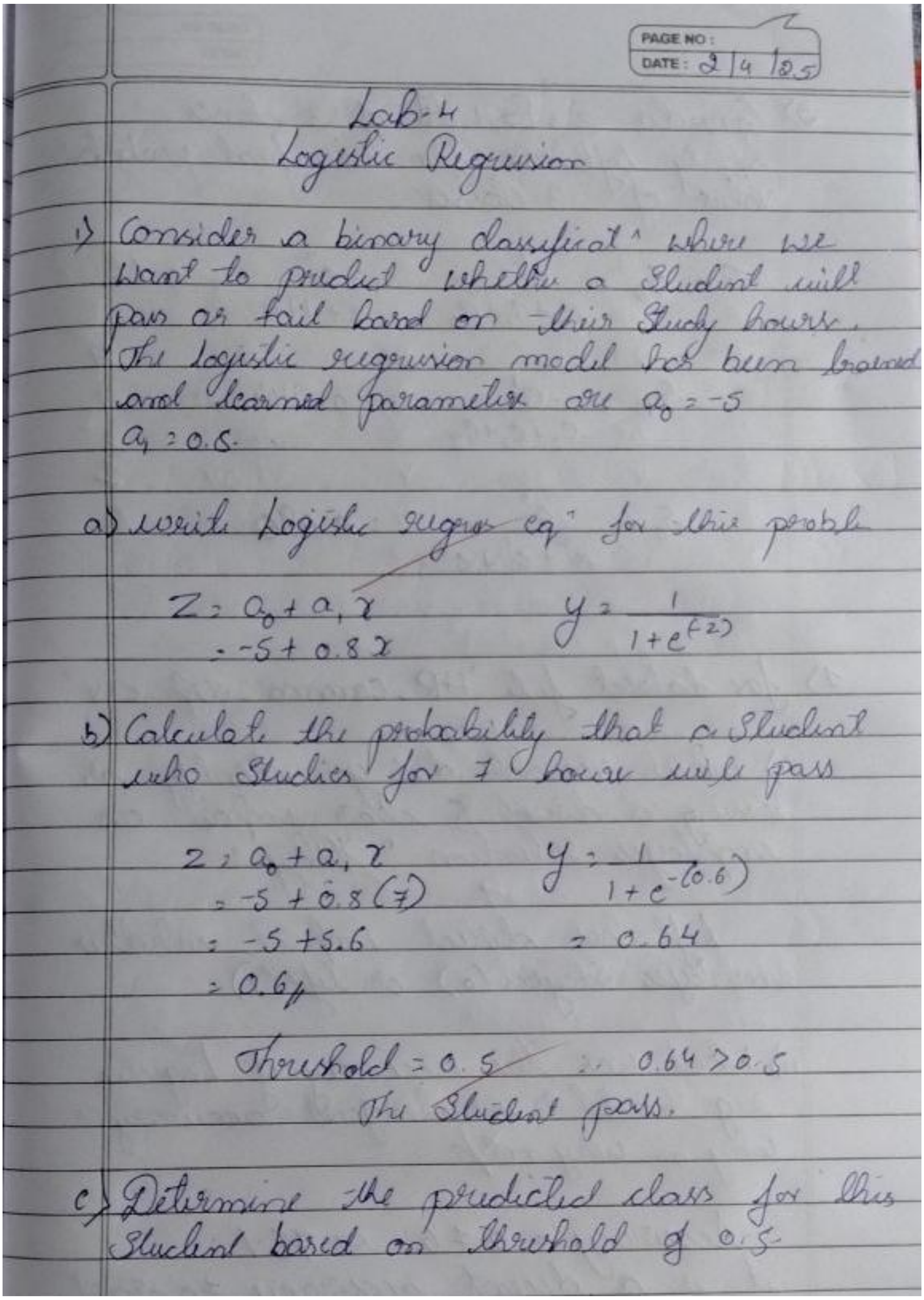
LAB-4

Logistic Regression



PAGE NO: 2
DATE: 2/4/25

Lab-4
Logistic Regression

1) Consider a binary classificat^n where we want to predict whether a Student will pass or fail based on their Study hours. The logistic regression model has been trained and learned parameter are $a_0 = -5$ $a_1 = 0.8$.

a) write Logistic regres eq^n for this problem

$$Z = a_0 + a_1 x \qquad y = \frac{1}{1+e^{(-z)}}$$
$$= -5 + 0.8x$$

b) Calculate the probability that a Student who Studies for 7 hours will pass

$$z = a_0 + a_1 x \qquad y = \frac{1}{1+e^{-(0.6)}}$$
$$= -5 + 0.8(7) \qquad = 0.64$$
$$= -5 + 5.6$$
$$= 0.6$$

Threshold = 0.5 ∴ 0.64 > 0.5
The Student pass.

c) Determine the predicted class for this Student based on threshold of 0.5

2) Consider z = [2, 1, 0] of 3 classes. Apply softmax fun' to find probability value of 3 classes

$$S = S_2 = \frac{e^2}{e_2+e_1+e_0} = \frac{e^2}{2+1+0} = 0.66$$

$$S_1 = \frac{e^1}{e_2+e_1+e_2} = 0.244$$

$$S_0 = \frac{e^0}{e^2+e^1+e^0} = 0.090$$

1) for dataset file 'HR_comma_sep.csv'

i) which variable did you identify as having a direct & clear impact on employee retention ? Why ?

Left has direct impact whether employee Stayed (0) or left (1)

ii) What was the accuracy of Logistic reg model ? is it good accuracy ? why or why not ?

Accuracy = 0.76 ≈ 76%
it is a decent accuracy 70-80%
if it is more than 80% then it is good accuracy.

2) for Zoo dataset

i) Did you perform any data preprocessing Steps? if yes, what are they and why they necessary.

Yes, we use label encoder for Converting categorical encoder data into numerical form.

It is necessary for further classification

ii) Were there any missing or inconsistent values in the dataset: how do handle them?

There are inconsistent values in the Animal names column as There is not Standardizat° of formats & units

iii) what does the confusion matrix tell you about the performance of your model?

all the No are in diagonal therefore No misidentificat° of classes.

iv) No class types are misclassified

PAGE NO :
DATE : 2/4/25

Lab-4
KNN (K- nearest neighbour)

Person	Age	Salary	Target	distance	Rank
A	18	50	N	$\sqrt{914}$	5
B	23	55	N	$\sqrt{2169}$	4
C	24	70	N	$\sqrt{1021}$	2
D	41	60	Y	40.44	3
E	43	70	Y	31.04	1
F	38	40	Y	60.07	6
X	35	100	?		

$$\rightarrow \sqrt{(18-23)^2+(50-55)^2} \rightarrow$$
$$= \sqrt{3^2+60^2}$$
$$= \sqrt{9+3600} = 60.07$$

$$\rightarrow \sqrt{(43-35)^2+(70-40)^2}$$
$$= \sqrt{8^2+30^2}$$
$$= \sqrt{964} = 31.04$$

$$\rightarrow \sqrt{(41-35)^2+(60-100)^2}$$
$$= \sqrt{6^2+40^2} = 40.44$$

K = 3 → N Y ~~N~~ Y
= Y

PAGE NO :
DATE :

For the Iris dataset:

Testin:

→ how to choose K value?

testing multiple K values & comparing their accuracy. The accuracy & error rate for K are compared and the most optional K is selected

here K = 3

diabetis dataset

⇒ what is the purpose of feature Scaling? how to perform it?

feature Scaling ensures all features contribute equally to the nearest neighbours Scaling is done so that feature like glucose / ~~gay~~ age don't dominate the one with the Smaller ranges.

2-4-25