بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

International Islamic University Chittagong

Department of Computer Science and Engineering

COURSE CODE : CSE-4878

COURSE TITLE : Machine learning & Data Mining Sessional

DATE OF SUBMISSION: 22/02/2022

SUBMITTED BY -

NAME	ID NO
SANJIDA ISLAM NOWRIN	C181223
SABRINA AFROZ	C181212
JANNATUN NAYEM HANI	C181229
FATEMA TUZ JOHORA	C181238

SUBMITTED TO -

Mrs. Subrina Akter
Assistant Professor
Department of CSE,
International Islamic University Chittagong.

SIGN: ______________________________ *MARKS*: __________________________

1. What is Regression?

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Regression helps investment and financial managers to value assets and understand the relationships between variables, such as commodity prices and the stocks of businesses dealing in those commodities.

2. Dataset Name: Hepatitis

3. Dataset Information:

Data Set Characteristics	**:** Multivariate
Attribute Characteristics	**:** Categorical, Integer, Real
Number of Instance	**:** 155
Number of Attributes	**:** 19
Number of class	**:** 1
Missing values?	**:** Yes
Area	**:** Life
Associated Tasks	**:** Classification

4. Input Attributes:

- age
- sex
- steroid
- antivirals
- fatigue
- malaise
- anorexia
- liver_big
- liver_firm
- spleen_palpable
- spiders
- ascites
- varices
- bilirubin
- alk_phosphate
- sgot
- albumin
- protime
- histology

5. Attribute Information:

Sl No.	Attribute/class Name	Type	Mean	Median	Mode		Unique Value
1	age	Continuous	42.1	39	30	0	10, 20, 30, 40, 50, 60, 70, 80
2	sex	Binary			female	0	Male, Female
3	steroid	Binary			True	1	True, False
4	antivirals	Binary			False	0	True, False
5	fatigue	Binary			True	1	True, False
6	malaise	Binary			False	1	True, False
7	anorexia	Binary			False	1	True, False
8	liver_big	Binary			True	10	True, False
9	liver_firm	Binary			False	11	True, False
10	spleen_palpable	Binary			False	5	True, False
11	spiders	Binary			False	5	True, False
12	ascites	Binary			False	5	True, False
13	varices	Binary			False	5	True, False
14	bilirubin	Continuous	1.4275	1	1	6	0.39,0.80,1.20,2.00, 3.00,4.00
15	alk_phosphate	Continuous	105.3254	85	85	29	33,80,120,160, 200, 250
16	sgot	Continuous	85.8940	59	20	4	13,100,200,300, 400, 500
17	albumin	Continuous	3.8173	4	4	16	2.1,3.0,3.8,4.5, 5.0, 6.0
18	protime	Continuous	61.85	100	100	67	10,20,30,40,50, 60,70,80,90
19	histology	Binary			False	0	True, False
20	Class	Binary			live	0	Live, Die

6. The Process of Converting Categorical Data into Numerical Data:

We used the following library function to convert categorical values into numerical values :

```
from sklearn.preprocessing import LabelEncoder
enc = LabelEncoder()
enc.fit(fl['Column-name'])
fl['Column-name'] = enc.transform(fl['Column-name'])
```

7. Linear Regression with Build in Function:

We have used the following build in functions for creating the linear regression model:

LinearRegression(), fit(), and **predict()**
from **sklearn.linear_model** package

8. Multiple Linear Regression with Raw Code (Using Matrix):

We followed the following steps for creating the multiple linear regression by using raw code:

Step 1: Import the required libraries.

Step 2: Read the file.

Step 3: Print the first five rows of dataset.

Step 4: Print column names.

Step 5: Read the column data into variables.

Step 6: Shape of our variables

Step 7: Plot the data on scatter plot.

Step 8: Convert our variables datatype from series to array.

Step 9: Number of rows in our dataset.

Step 10: Create a “ones” matrix.

Step 11: Reshape our data so that we can perform operations like addition and multiplication with x_bias.

Step 12: Create a major matrix with all the columns like x_bias.

Step 13: Print the major matrix.

Step 14: Find transpose of a matrix.

Step 15: Perform multiplication.

Step 16: Find inverse.

Step 17: perform multiplication.

Step 18: Finding coefficients.

Step 19: print the coefficient values.

Step 20: Predict the values based on the calculated coefficient values.

Summary: We have used **LinearRegression(), fit(),** and **predict()** from **sklearn.linear_model** package which reduces steps for multiple linear regression. This is the benefit of using built-in functions. But in case of using raw code, there are many steps which makes the whole process lengthy and complicated, though raw code is the best practice to understand how the algorithm actually works. Built-in function makes the process handy for us by allowing more times to focus on other techniques.

==============x============