A.

* Load the Cleaveland Heart disease dataset: [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/45/heart+disease)

**Cleveland Heart-disease dataset**

Attribute Information:

1. Age (in years)

2. Sex (1 = male; 0 = female)

3. cp -chest pain type

4. trestbps - resting blood pressure (anything above 130-140 is typically cause for concern)

5. chol-serum cholestoral in mg/dl (above 200 is cause for concern)

6. fbs - fasting blood sugar ( > 120 mg/dl) (1 = true; 0 = false)

7. restecg - resting electrocardiographic results (0 = normal;1 = having ST-T waveabnormality; 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)

8. thalach-maximum heart rate achieved

9. exang - exercise induced angina (1 = yes; 0 = no)

10. oldpeak - depression induced by exercise relative to rest

11. slope - slope of the peak exercise ST segment (1 = upsloping; 2 = flat Value; 3 =downsloping)

12. ca - number of major vessels (0-3) colored by flourosopy

13. thal - (3 = normal; 6 = fixed defect; 7 = reversable defect

14. **num** (target) - diagnosis of heart disease (angiographic disease status)( 0: < 50% diameternarrowing ; 1: > 50% diameter narrowing)

* check the type of data variable
* Display last five rows of the dataset
* Experiment with the database by attempting to distinguish presence (values 1,2,3,4) from absence (value 0)
* Change instances with labels 2,3 and to 1.
* The feature 'ca' has missing values that are given as '?'. Let us replace the '?' with nan and then fill those missing values using 'mean' imputation strategy.
* Remove the target variable from heart\_data
* Draw a heatmap to understand the correlation between Input features
* Split the data for training and testing at 80:20
* Normalizing features for training using Standardscaler
* Perform Classification using logistic regression and calculate the training score
* Prepare a confusion matrix and classification report (accuracy, precision etc) for the same

Challenge to get additional points

* Can you do a Hyperparameter tuning of the logistic regression model with RandomizedSearchCV and GridSearchCV?

B.

Apply LASSO regression on the attached dataset.

About Dataset

Hitters Case Study

Major League Baseball Data from the 1986 and 1987 seasons.

This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. This is part of the data that was used in the 1988 ASA Graphics Section Poster Session. The salary data were originally from Sports Illustrated, April 20, 1987. The 1986 and career statistics were obtained from The 1987 Baseball Encyclopedia Update published by Collier Books, Macmillan Publishing Company, New York.

What are the datatypes?

DESCRIBE AND TRY TO FILL NA VALUES

Apply one hot encoding and detect the outliers

Apply OLS, ridge and lasso regression

Which model gives the highest accuracy?