Assignment-4

Problem Statement 1: The data (sample) were collected in São Paulo — Brazil, in a university, where there are some parties with groups of students from 18 to 28 years of age (average). The dataset used for this activity has 7 attributes, being a Target, with a period of one year. You have to predict the quantity of beer consumption based on the features that contain climate conditions.

Dataset Description:

I. Data: date of the record
II. Temperatura Media (C): Average temperature of the day in celsius
III. Temperatura Minima (C): Minimum temperature of the day in celsius
IV. Temperatura Maxima (C): Maximum temperature of the day in celsius
V. Precipitacao (mm): Percipitation in mm
VI. Final de Semana: If the day is the weekend or not
VII. Consumo de cerveja (litros): Beer consumption in liters

Write a Python code to perform the following tasks mentioned:

1. Load the dataset, check its shape
2. Rectify the data of the first four columns
   Hint: Check columns 'Temperatura Media (C)', 'Temperatura Minima (C)', 'Temperatura Maxima (C)', and 'Precipitac'
   Fix the following errors present in these features
3. Create new features using the 'Data' feature and the make 'Data' column as index
   Hint: Create a new feature 'Month' from the dates, consisting of the month of the year.
   Create a new feature 'Day' from the dates, consisting of the day of the week.
   Set values from the 'Data' column as indexes. Use code snippet:
   `df1.set_index('Data',inplace=True)`
4. Handle null and duplicate values
5. Check the data type of the features and convert them to the appropriate data type
6. Analyze features with outlier values
7. Plot and analyze the correlation
8. Split the dataset for training and testing
9. Train a linear regression model and print the intercept and coefficients
10. Evaluate the model using the R2 score, mean absolute error, and root mean squared error

Problem Statement 2: You are provided with the California housing dataset. Based on the given parameters of a house, predict its price.

Dataset Description:

The dataset contains nine features:

I. longitude: A measure of how far west a house is; a higher value is farther west
II. latitude: A measure of how far north a house is; a higher value is farther north
III. housingMedianAge: Median age of a house within a block; a lower number is a newer building
IV. total rooms: Total number of rooms within a block
V. total bedrooms: Total number of bedrooms within a block
VI. population: Total number of people residing within a block
VII. households: Total number of households, a group of people residing within a home unit, for a block
VIII. median income: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
IX. median house value: Median house value for households within a block (measured in US Dollars)

Write a Python code to perform the following tasks mentioned:

1. Load the data, check its shape and check for null values
2. Split the dataset for training and testing - 1000 instances for testing
3. Train the model using sklearn (Apply linear regression to train a model for prediction)
4. Predict the prices on test data and evaluate the model by r2 score and mean absolute error
5. Find coefficient and intercept using the trained model

Problem Statement 3: You are provided with the medical cost dataset. You need to predict individual medical costs billed by health insurance.

Dataset Description:

I. age: age of the primary beneficiary
II. sex: gender of primary beneficiary female, male
III. bmi: Body mass index, providing an understanding of the body, weights that are relatively high or low relative to height, an objective index of body
IV. weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
V. children: Number of children covered by health insurance / Number of dependents
VI. smoker: Smokes or not
VII. region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest
VIII. charges: Individual medical costs billed by health insurance

Write a Python code to perform the following tasks mentioned:

1. Load the data, check its shape and check for null values
2. Convert categorical features to numerical values (Use One-Hot Encoding)
3. Split the dataset for training and testing
4. Train the model using sklearn – Linear Regression
5. Find the intercept and coefficient from the trained model
6. Predict the prices of test data and evaluate the model using calculated r2 score and root mean squared error

Problem Statement 4: You are provided with the '50_Startups' data. Using the given features, you must predict the profit of these startups.

Dataset Description:

I. R&D Spend: Expenditures in Research and Development
II. Administration: Expenditures in Administration
III. Marketing Spend: Expenditures in Marketing
IV. State: In which state the company belongs to
V. Profit: The profit made by the company

Write a Python code to perform the following tasks mentioned:

1. Load the data, check its shape and check for null values
2. Convert categorical features to numerical values using Label Encoder
3. Split the dataset for training and testing
4. Train the model using sklearn (linear regression), also find the intercept and coefficient from the trained model
5. Predict the profits of test data and evaluate the model using r2 score and mean squared error
6. Regularize the model using Ridge Regression and find the Score
7. Regularize the model using Lasso Regression and find the Score