**Department of Computer Science**

**MAI272 – Advanced Machine Learning**

**Program\_1: Simple Linear Regression:**

Each Sample in the dataset describes a house in Boston town. The data was drawn from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. The attributes are deﬁned as follows

* Crime\_Rate per capita crime rate by town
* Zone\_proportion of residential land zoned for lots over 25,000 sq.ft.
* Non\_Retail\_Acres proportion of non-retail business acres per town
* CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
* NOX nitric oxides concentration (parts per 10 million)
* Avg\_rooms average number of rooms per dwelling
* AGE proportion of owner-occupied units built prior to 1940
* Distance weighted distances to five Boston employment centres
* RAD index of accessibility to radial highways
* Property\_TaxRate full-value property-tax rate per 10 000 USD
* PTRATIO pupil-teacher ratio by town
* B 1000 (Bk - 0.63)^2 where Bk is the proportion of black people by town
* LSTAT % lower status of the population
* MValue Median value of owner-occupied homes in $1000's

**Use the dataset and perform the following activities:**

1.    Consider the column, ‘Avg\_room’, as predictor, and ‘MValue’ as the target variable

2.    Visualize the association between the predictor and the target using scatter plot.

3.    Split the data into train and test datasets, in the ratio of 70:30.

4.    Build a Linear Regression model using training dataset, to predict the target variable.

5.    Observe the coefficient and intercept values for the model.

6.    Evaluate the model using mean squared error values and R-squared values on the training

and the testing datasets.

7. Provide suitable inference on the model created.

*Note: Dataset can be pre-processed if required*

**Evaluation Rubrics**

|  |  |
| --- | --- |
| **Content** | **Score** |
| Explanation / Equation / Worked-out examples for demonstrating the flow of the algorithm | 3 Marks |
| Use of appropriate functions / parameters | 5 Marks |
| Concept Clarity (Algo & Dataset) | 2 Marks |
| Total | 10 Marks |

**Department of Computer Science**

**MAI272 – Advanced Machine Learning**

**Program\_2: Multiple Linear Regression**

Each Sample in the dataset describes a house in Boston town. The data was drawn from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. The attributes are deﬁned as follows

* Crime\_Rate per capita crime rate by town
* Zone\_proportion of residential land zoned for lots over 25,000 sq.ft.
* Non\_Retail\_Acres proportion of non-retail business acres per town
* CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
* NOX nitric oxides concentration (parts per 10 million)
* Avg\_rooms average number of rooms per dwelling
* AGE proportion of owner-occupied units built prior to 1940
* Distance weighted distances to five Boston employment centres
* RAD index of accessibility to radial highways
* Property\_TaxRate full-value property-tax rate per 10 000 USD
* PTRATIO pupil-teacher ratio by town
* B 1000 (Bk - 0.63)^2 where Bk is the proportion of black people by town
* LSTAT % lower status of the population
* MValue Median value of owner-occupied homes in $1000's

**Use the dataset and perform the following activities:**

1. Consider the attributes Av\_room, distance, Property\_TaxRate, Non\_Retail\_Acres as predictors and MValue as the target variable.
2. Calculate the Variance Inflation Factor for each of the selected predictors. Based on the VIF factor, finalize the list of predictors
3. For every predictor identified, visualize its association with the target column using scatter plot.
4. Split the data into train and test datasets, in the ratio of 67:33
5. Build a multiple linear regression model to predict the target variable using the selected predictors
6. Observe the coefficients and intercept values for the model
7. Evaluate the model using mean squared error values, R-squared Values and adjusted R-squared values on the train and test dataset
8. Provide suitable inference on the model created.

*Note: Dataset can be pre-processed if required*

**Evaluation Rubrics**

|  |  |
| --- | --- |
| **Content** | **Score** |
| Explanation / Equation / Worked-out examples for demonstrating the flow of the algorithm | 3 Marks |
| Use of appropriate functions / parameters | 5 Marks |
| Concept Clarity (Algo & Dataset) | 2 Marks |
| Total | 10 Marks |

**Department of Computer Science**

**MAI272 – Advanced Machine Learning**

**Program\_3a: Logistic Regression:**

Perform the following using Program 3a\_dataset

1. Consider duration, age, and campaign columns as predictors and the column Y as target  variable. Y says whether the client will subscribe a term deposit or not.
2. Use 70% of the data as training data set and 30% of data as testing data set.
3. Build a logistic regression model.
4. Determine the classification accuracy rate of the model on both train and test set.
5. Display Classification Report and Confusion Matrix
6. Provide suitable inference on the model created.

**Program\_3b: Logistic Regression:**

Perform the following using Program3b\_dataset

1. Consider 'variance', 'skewness', 'curtosis', 'entropy' columns as predictors and the column 'class' as the target  variable
2. Build a logistic regression model.
3. Calculate the classification accuracy, precision and recall for the model
4. Display Classification Report and Confusion Matrix
5. Provide suitable inference on the model created.

*Note: Dataset can be pre-processed if required*

**Evaluation Rubrics**

|  |  |
| --- | --- |
| **Content** | **Score** |
| Explanation / Equation / Worked-out examples for demonstrating the flow of the algorithm | 3 Marks |
| Use of appropriate functions / parameters | 5 Marks |
| Concept Clarity (Algo & Dataset) | 2 Marks |
| Total | 10 Marks |

**Department of Computer Science**

**MAI272 – Advanced Machine Learning**

**Program\_4a: Polynomial Regression:**

**Use the program\_4\_dataset attached and perform the mentioned activities.**

**Dataset description:**

The dataset contains outdoor temperature and pressure data between the duration as given in the timestamp. The dataset is presented in OS\_time, temperature, pressure humidity and actual date and time format. Interesting patterns between temperatures, pressure can be found for understanding the humidity or rainfall during the given period. This can also be used to understand environmental changes over time.

1. Consider Temperature as predictor and the target as pressure
2. Build a model using polynomial regression with at least 6 different degrees with 3 different train-test split as 66-44, 70-30, 80-20.
3. Calculate R2 square. Mean Squared Error, Mean Absolute Error and validate your model
4. Provide suitable inferences on the model created.

*Note: Dataset can be pre-processed if required*

**Evaluation Rubrics**

|  |  |
| --- | --- |
| **Content** | **Score** |
| Explanation / Equation / Worked-out examples for demonstrating the flow of the algorithm | 3 Marks |
| Use of appropriate functions / parameters | 5 Marks |
| Concept Clarity (Algo & Dataset) | 2 Marks |
| Total | 10 Marks |

**Department of Computer Science**

**MAI272 – Advanced Machine Learning**

**Program\_5: Lasso and Ridge Regression:**

**Dataset description:**

All attributes except Category and Sex are numerical. The laboratory data are the attributes 5-14.

1) X (Patient ID/No.)

2) Category (diagnosis) (values: '0=Blood Donor', '0s=suspect Blood Donor', '1=Hepatitis', '2=Fibrosis', '3=Cirrhosis')

3) Age (in years)

4) Sex (f,m)

5) ALB

6) ALP

7) ALT

8) AST

9) BIL

10) CHE

11) CHOL

12) CREA

13) GGT

14) PROT

**Perform the following using Program\_5\_dataset**

* 1. Build a simple linear Regression considering ‘CREA’ as predictor variable and ‘PROT’ as Target variable (for all models)
  2. Calculate the VIF for all the attributes except X (Patient ID/No.), 2) Category 3) Age, and 4) Sex. Based on the VIF factor, finalize the list of predictors
  3. Split the data into train and test datasets in the ratio of 80:20 (for all experiments)
  4. Build a multiple linear regression model to predict the target variable ‘PROT’ using the selected predictors
  5. Build a polynomial regression model using the attribute that has highest VIF for 3 different degrees
  6. Build a Ridge regression model using the attribute that has highest VIF for 13 different values of alpha ranging from 0.1 to 150
  7. Build a LASSO regression model using the attribute that has highest VIF for different values of alpha ranging from 0.1 to 100
  8. Use N-fold cross validation for training the model created.
  9. Evaluate the models using mean squared error values, R-squared Values and adjusted R-squared values (wherever applicable) on the train and test dataset
  10. Print intercept and coefficient values wherever appicable
  11. Provide a comparative analysis graph of models built and give conceptual inferences on all the models built

*Note: Dataset can be pre-processed if required*

**Evaluation Rubrics**

|  |  |
| --- | --- |
| **Content** | **Score** |
| Explanation / Equation / Worked-out examples for demonstrating the flow of the algorithm | 3 Marks |
| Use of appropriate functions / parameters | 5 Marks |
| Concept Clarity (Algo & Dataset) | 2 Marks |
| Total | 10 Marks |