**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 |