Learning outcomes from the project

***Task2: Dataset- Team***

In this project we are going to make a machine learning linear regression model by using python language. We are going to predict some values from the data which is already given to us.

Steps included in this process are:

⁃ Forming hypothesis :

Form a hypothesis means, to give answer to your research question that has not yet been tested. It will lead to more efficient and effective research outcomes. By forming hypothesis values can be predicted easily. In this project we are going to predict how many medals a country will win in olympics.

⁃ Data interpretation :

Interpreting data is most important for knowing the format and nature of data

Here,

• First column indicates which country is participating

• Second column indicates which year the olympics were held

• Third column indicates how many athlete participated from one country

• Fourth column indicates how many previous medals they won

• Last column is the predicted value of medals that were won by different teams

⁃ Data reshaping :

Data should be reshaped in order to predict values so that it can be formatted in a way that the network is understood and can be used for training. In this case data is already reshaped as we have two separate columns to compare where predictors and target values are in a single row. Here, we will use “Athletes” column and “Pre Medal” column to predict how many medals a team would win in a year.

⁃ Cleaning data :

the process of fixing or removing incorrect, corrupted, missing, duplicate, or incomplete data within a dataset. It is important to improve overall productivity and decision making during prediction. In this case there are some missing values that needs to be handled. Here, in the column “Pre Medal” there are missing values of those teams who didn’t played the olympics in some year. We got to clean data in order to handle missing values

⁃ Error metrices :

It is used to evaluate the performance of our model. Predicted value would be different from the actual value. We have to figure out whether our predictions are good or bad. We will use Mean Absolute Error. The difference of actual number of medals to the predicted number of medals then the modulus of it is called error. Mean Absolute Error is the sum of error divided by the number of terms. By finding the magnitude of error we could know about the accuracy of the model.

⁃ Split data :

In this process we split the data in two parts first is Training data and second in Testing data

• Training data- It is an input dataset used to train a machine learning model. These models use training data to learn and refine rules to make predictions on unseen data points.

• Test data- It is a set of data used to validate the correctness, completeness, and quality of a software program or system.

We split data in training set and testing set to understand how good the model we have trained is, so that we can decide if we want to deploy the model

Without a dedicated testing set, the risk of overfitting increases when a model adapts too closely to the training data.

⁃ Methods to train data:

We will use linear regression to train our model

Linear regression model is a model that compares two variables by plotting graph where all the points lies near the straight line. Basic formula of regression line is

Y=b0+b1x1

It will tell the variation among the variables. This helps to predict new data with the help of old data

Here, we have two predictors

1. Number of athletes

2. Number of medal

So, we will use this formula

Y=b0+b1x1+b2x2

By using linear regression we will train the data and predict the new values with the help of previous values.