**The Machine Learning and Deep Learning Show**

**Assignment 01**

Submission Date: 27 May 2021

**Name: Sriniketh J Participant ID: 22**

**Task 1: Dataset**

Link**:** [**Dataset**](https://archive.org/download/ages-and-heights/AgesAndHeights.pkl)

**Task 2: Coding**

**Drive:** [**Drive Folder**](https://drive.google.com/drive/folders/1X9FgcARAGW1tBFFptW8LqSHd8Trv-1Kf?usp=sharing)

**Github:** [**Repository Link**](https://github.com/srini047/MLDLS-2021)

**Introduction**

Hello Everyone for this practical Introduction to Machine Learning using Simple Linear Regression. So, let’s get started:

So, let’s get familiarised with the terms to be used:

**Machine Learning (ML)**: ML is an application of Artificial Intelligence (AI) that provides systems the ability to automatically learn themselves and improve from the experience without being explicitly programmed. ML focuses on the development of computer programs that can access data and use it to learn themselves.

**Data Set**: A collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer.

**Data Visualisation**: It is a representation of data or information in a graph, chart, or other visual formats which is helpful to conduct analyses such as predictive analysis which can serve as helpful Visualisation to present.

**Data Cleaning**: It is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

**Supervised Learning**:  The model is trained using ‘labeled data’. Datasets are said to contain labels that contain both input and output parameters. To simplify – ‘Data is already tagged with the correct answer’.

**Simple Linear Regression**: It is a Regression Model that estimates the relationship between the independent variable and the dependent variable using a straight line [y = mx + c], where both the variables should be quantitative.

**Models**: Those are output by algorithms and are comprised of model data and a prediction algorithm.

**Training Model**: In supervised learning, an ML Algorithm builds a model by examining many examples and attempting to find a model that minimizes loss and improves prediction accuracy.

These are the few terms used and to get familiarized with. Now let’s get started with the analysis and prediction of the model. I am going to use supervised data and simple linear regression for analysis and prediction. The Ultimate goal is the predict the height of a person provides his age using the trained model to the highest achievable accuracy using available data. Where, age is the dependent variable and height is the independent variable. I have used the universal favourite programming language for ML i.e. **Python** to build and train the ML model and Google Colab Environment.

**DISCLAIMER**: Since, the dataset has values until 18 years this model works well the most within that age limit.

***The steps involved are:***

1. Importing the dataset.

2. Visualising the Data

3. Data Cleaning

4. Build the Model and Train it

5. Make Predictions on Unseen Data

**1. Importing Data Set:**

The first and foremost thing we need to do is import the dataset. We have various websites which have these datasets to be used by anyone. So let’s get started by importing the dataset.

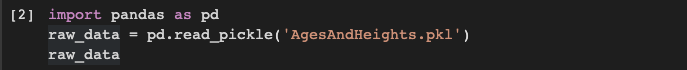
Simple Linear Regression - Import data 

This single line of code helps us fetch the data used for the prediction from the URL directly.

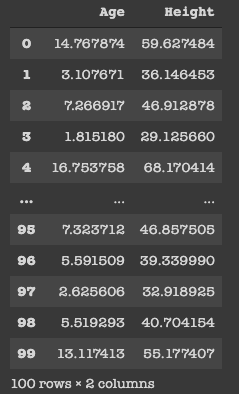
[Dataset](https://archive.org/download/ages-and-heights/AgesAndHeights.pkl) <- Click the link to fetch the dataset which is the above-mentioned URL.

**2. Visualising the Data:**

In this step after importing the data and mounting it with Colab let’s have an overview of the dataset by importing a Module called pandas. Since the dataset we have has an extension of .pkl we just view it by the function available in the pandas library.



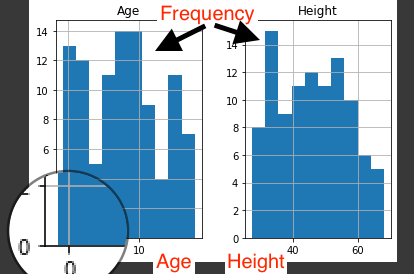
We import the library to read the dataset and store it in a variable called **raw\_data.**We then display the content of **raw\_data** which is in a tabulated format.



We can see the data which we have and it contains only 2 columns namely, Age (in years) and Height (in inches) and 100 rows which is actually the representation of a person.

history

This single line of code has a great impact on the way we look at the dataset. We only had a numerical view of the dataset but we can now run this cell to get a histogram view of the dataset which is very helpful. It represents the data present in the individual columns as individual graphs.



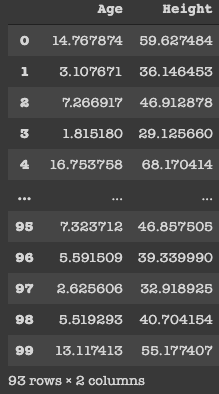
The Y-axis in both the plots refers to frequency and X-axis represents Age and Height respectively.

**3. Data Cleaning:**

We have to build the model using valid datasets and clean the unaccountable Data. In the above image, we can know that there are a few entries that have an age less than zero which is meaningless. Hence, we need to clean those data to get better accuracy.

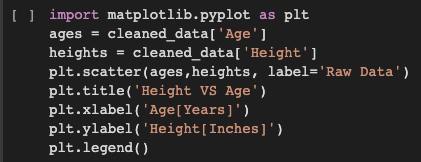
Data Cleaning

I use variable **cleaned\_data** to store the valid age values and display them to the user.

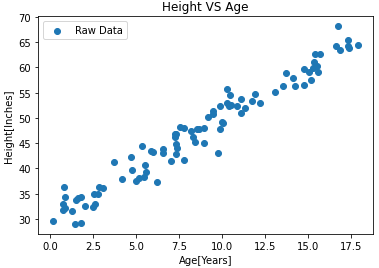


Initially, we had 100 rows but after performing Data Cleaning it’s pretty clear that there are seven rows which we had age < 0 and we have removed them. As a professional, we aren’t supposed to delete the data as we are reducing the data and thereby accuracy of our model gets reduced. To keep it simple I have just removed them.

Visualize the Cleaned Data: I have now used the cleaned data and visualized it in the form of a graph.

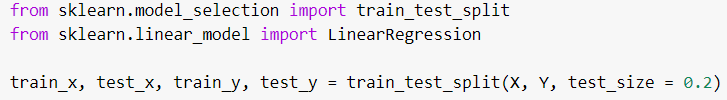


To plot graphs in python I import matplotlib.pyplot library. I represent Age on X-axis and Height on Y-axis. The points in the plot refer to the Raw data.



**4. Build the Model and Train it:**

To build the model I have used the Linear Regression Algorithm, which is the most fundamental algorithm in the entire ML world. For which I have used the scikit-learn module which has the linear-model library which contains LinearRegression module.

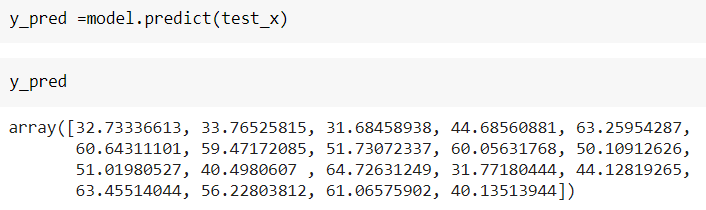


I have splitted the entire dataset as 80% -> Train and 20% -> Test and saved them in the variables train\_x, test\_x, train\_y and test\_y. So if we see the shape of train\_x, test\_y we find that 17 rows are for train and 19 rows are for test. I have splitted thuis for the cleaned data.

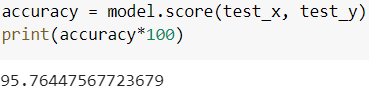
The LinearRegression function can be called by assigning a name like ‘model’ so that it is easier to access it. Then I have fitted the model using the train\_x and train\_y as parameters.



Then for every value in test\_x, I find the value of y using the model. Then I store that in a variable called as y\_pred.



Then to find the accuracy of the model I use the score function which predicts how accurate our model is for every value in test\_x, the value we get for test\_y. I use the score function and store the accuracy in the variable ‘accuracy’.



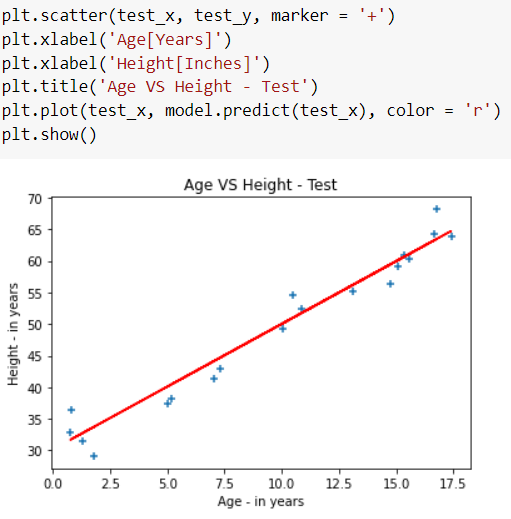
From this we can see that the accuracy of the model is around 95%. So, we can conclude that our model has worked pretty well.

To make this even more lively not just depending on the mathematical stats, we can plot a graph between age and height for train and test data so that it brings entirety to the conclusion.

**For Train Set:**

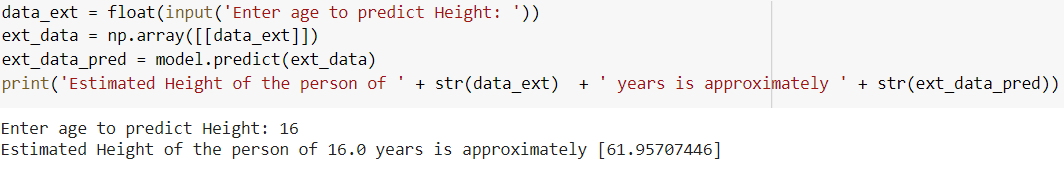


**For Test Set:**



**5. Prediction on External Data:**

With the help of this trained model, we can now make accurate predictions.



So, we can see for any given age we find the possible height in inches. Finally, we have successfully trained the model and verified it with help of visualizations and predictions on unseen data.