Third Year B. Tech (EL & CE)

Semester: VI **Subject:** Data Science for Engineering

Name: Shreerang Mhatre **Class: TY**

Roll No: 52 **Batch: A2**

Experiment No: 05

Name of the Experiment: **Data Science Fundamentals**

Performed on: 23/03/2024

Submitted on: 18/04/2024

Problem Statement:

Write a python program to predict the height of a person providing his age using the trained model to the highest achievable accuracy using available data.

Perform following steps:

i. Importing the dataset. Link of Data.

ii. Perform exploratory analysis of the data: Print features, Shape, Size, labels, head records, data types, outliers etc.

iii. Data Cleaning.

iv. Build the Model and Train it.

v. Make Predictions on Unseen Data.

Analyze the performance of the model.

localhost:8888/notebooks/DSE%20Practicals/EXP%20-%205/Shreerang%20Mhatre_52_DSE_exp5.ipynb

jupyter Shreerang Mhatre_52_DSE_exp5 Last Checkpoint: 3 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

In [1]:
```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
```

In [2]:
```
# Load data
data_url = 'https://archive.org/download/ages-and-heights/AgesAndHeights.pkl'
raw_data = pd.read_pickle(data_url)
```

In [3]:
```
# Perform exploratory analysis
print("Features:")
print(raw_data.columns)
print("\nShape:")
print(raw_data.shape)
print("\nSize:")
print(raw_data.size)
print("\nLabels:")
print(raw_data['Height'].unique())
print("\nHead Records:")
print(raw_data.head())
print("\nData Types:")
print(raw_data.dtypes)
print("\nOutliers:")
print(raw_data[raw_data['Age'] <= 0])
```

```
Features:
Index(['Age', 'Height'], dtype='object')

Shape:
(100, 2)
```

localhost:8888/notebooks/DSE%20Practicals/EXP%20-%205/Shreerang%20Mhatre_52_DSE_exp5.ipynb

jupyter Shreerang Mhatre_52_DSE_exp5 Last Checkpoint: 3 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

```
Features:
Index(['Age', 'Height'], dtype='object')

Shape:
(100, 2)

Size:
200

Labels:
[59.62748426 36.14645274 46.91287846 29.12565975 68.1704139  44.34636503
 47.71361295 37.82994409 44.79306512 34.05901466 46.14234545 56.45678228
 59.87003819 28.6478337  64.23638736 34.24599643 45.29759191 54.75189183
 32.93518571 51.99837    29.88392844 31.52350033 55.70481247 63.86109721
 52.9199051  32.83092177 37.36102444 60.23327195 35.01026126 59.12701931
 31.96080741 48.21846885 32.56232215 47.75274969 36.24179872 33.78175852
 27.33251348 38.2877098  59.10262494 43.28700254 39.66542429 44.95296111
 37.43903005 52.31971537 38.09928941 29.50846768 32.17702291 42.91083707
 52.33957205 62.64153739 36.34620991 62.68458017 50.12874248 48.01159919
 41.30798526 31.82734682 43.07902442 43.03917945 44.13595398 30.80134635
 65.38850138 52.55874397 57.54336087 50.73552242 56.35839211 41.5764792
 32.30251117 42.33202065 49.22075916 53.45369077 47.80362461 34.93143275
 29.00769567 52.33693502 63.38905063 52.92067293 51.44539695 48.99409356
 47.33274976 54.52644907 46.21415637 50.94870845 34.24840688 47.99492935
 41.41827569 43.41211009 64.30784502 43.87201084 60.98925274 64.3862124
 56.36445828 27.81905232 57.97662043 53.81673813 58.9359501  46.85750545
 39.33998993 32.91892484 40.70415428 55.17740742]

Head Records:
         Age     Height
0  14.767874  59.627484
1   3.107671  36.146453
2   7.266917  46.912878
3   1.815180  29.125660
4  16.753758  68.170414
```

localhost:8888/notebooks/DSE%20Practicals/EXP%20-%205/Shreerang%20Mhatre_52_DSE_exp5.ipynb

jupyter Shreerang Mhatre_52_DSE_exp5 Last Checkpoint: 3 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

```
 47.33274976 54.52644907 46.21415637 50.94870845 34.24840688 47.99492935
 41.41827569 43.41211009 64.30784502 43.87201084 60.98925274 64.3862124
 56.36445828 27.81905232 57.97662043 53.81673813 58.9359501  46.85750545
 39.33998993 32.91892484 40.70415428 55.17740742]

Head Records:
        Age     Height
0  14.767874  59.627484
1   3.107671  36.146453
2   7.266917  46.912878
3   1.815180  29.125660
4  16.753758  68.170414

Data Types:
Age       float64
Height    float64
dtype: object

Outliers:
         Age     Height
13 -0.163532  28.647834
20 -0.683017  29.883928
25 -0.146392  32.830922
30 -0.780853  31.960807
36 -0.087958  27.332513
59 -0.548488  30.801346
91 -0.328780  27.819052
```

In [4]:
```
# Data cleaning
cleaned_data = raw_data[raw_data['Age'] > 0]
```

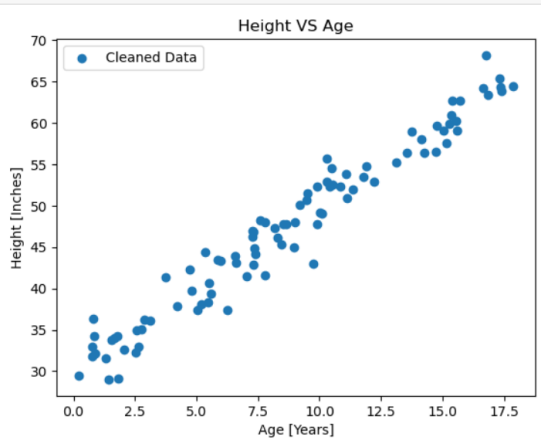
In [5]:
```
# Visualization
ages = cleaned_data['Age']
heights = cleaned_data['Height']
plt.scatter(ages, heights, label='Cleaned Data')
```

localhost:8888/notebooks/DSE%20Practicals/EXP%20-%205/Shreerang%20Mhatre_52_DSE_exp5.ipynb

jupyter Shreerang Mhatre_52_DSE_exp5 Last Checkpoint: 4 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

In [5]:
```
# Visualization
ages = cleaned_data['Age']
heights = cleaned_data['Height']
plt.scatter(ages, heights, label='Cleaned Data')
plt.title('Height VS Age')
plt.xlabel('Age [Years]')
plt.ylabel('Height [Inches]')
plt.legend()
plt.show()
```



localhost:8888/notebooks/DSE%20Practicals/EXP%20-%205/Shreerang%20Mhatre_52_DSE_exp5.ipynb

jupyter Shreerang Mhatre_52_DSE_exp5 Last Checkpoint: 4 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [6]:
```
# Prepare data for modeling
X = cleaned_data[['Age']]
y = cleaned_data['Height']
```

In [7]:
```
# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [8]:
```
# Model initialization and training
model = LinearRegression()
model.fit(X_train, y_train)
```

Out[8]:
```
LinearRegression()
```

In [9]:
```

# Predictions on test set
y_pred = model.predict(X_test)
```

In [10]:
```
# Model evaluation
mse = mean_squared_error(y_test, y_pred)
```

In [11]:
```
# Extracting coefficients and intercept
m = model.coef_[0]
c = model.intercept_
```

In [12]:
```
print("\nSlope (m):", m)
print("Intercept (c):", c)
print("Mean Squared Error:", mse)
```

```
Slope (m): 1.9766634632677893
Intercept (c): 30.122406680138038
Mean Squared Error: 3.870169406686577
```

MIT-WPU
MIT WORLD PEACE UNIVERSITY | PUNE

Exp 5

Shreerang Mhatre 52_A2

Post lab Questions

Q1) Explain the difference between linear regression & multiple linear regression.

→ ① Simple Linear Regression -

In this, there is only one independent variable (predictor variable) and one dependent variable (responsible variable). The goal is to model the relationship between these two variables using a linear equation, typically in the form of $y = mx + b$, where y is the dependent variable, x is independent variable & m is the slope.

② Multiple Linear Regression -

In multiple linear regression, there are multiple independent variables influencing a single dependent variable. The linear equation takes the form $y = b_0 + b_1x_1 + b_2x_2 + \cdots b_nx_n$ where y is the dependent variable $x_1, x_2, \cdots x_n$ are the independent variables, b_0 is the intercept & $b_1, b_2, \cdots b_n$ are the coefficients of independent variables.

www.mitwpu.edu.in

Q2) Describe different measures to analyze the errors in regression.

→ ① Residual Analysis -

Residuals are the differences between the observed value and the predicted value from the regression model. Analyzing residuals involves checking for patterns or trends in the residuals plot.

② R-squared (coefficient Determination)

R-squared measures the proportion of variation in the dependent variable that is explained by the independent variables in the regression model.

③ Adjusted R-squared -

Adjusted R-squared accounts for the number of independent variables in the model, providing a more accurate measure of goodness-of-fit for models with multiple predictors.

④ Mean Absolute Error (MAE) & RMSE

These metrics quantify the average magnitude of errors between & predicted values.

Q3) Differentiate & describe the properties of dependent & independent variables.

→ ① Dependent variable -

Also known as the response variable, the dependent variable is the outcome or target variable that we want to predict or explain based on the independent variable.

② Independent variable -

Also known as the predictor variable, the independent variable is the input or explanatory variable that is hypothesized to have an effect on the dependent variable.

③ Properties of Independent variable -

Independent variables should ideally have a linear relationship with the dependent variable, be independent of each other (no multicollinearity), be normally distributed, and have homoscedasticity. (constant variance.

Q4) Enlist five data science applications where regression can be used and justify why it is used.

→ ① Predictive Modelling -
Regression is widely used for predictive modelling tasks such as predicting sales revenue, stock prices, housing prices, or customer churn rates.

② Risk Assessment -
Regression can be used in risk assessment applications such as credit scoring, insurance premium estimation, or predicting the likelihood of loan defaults.

③ Demand Forecasting -
In industries like retail & supply chain management, regression is used for demand forecasting.

④ Healthcare Analytics -
Regression is applied in healthcare analytics for tasks such as predicting patient outcomes, estimating treatment effectiveness.

www.mitwpu.edu.in