A PROJECT REPORT ON

"SOCR-Height Weight Dataset"

Prem Kumar R 1JB18EC071

Shree Charan 1JB18EC087

INDEX

ABSTRACT

In 1993 a territory-wide cross-sectional growth survey on 25,000 Chinese children from birth to 18 years was performed in Hong Kong. Compared to the last growth survey in 1963, definite secular changes were observed. There was an increase of final adult standing height of 3.6 cm in boys and 2.7 cm in girls, in which 1.8 cm and 0.5 cm respectively for boys and girls was accounted for by the sitting height. Thus, most of the height increase had occurred in the leg length in girls, but in boys only half of it. The height difference was more marked during the pubertal years because secular change had brought about an earlier sexual maturation, including an advancement of median menarcheal age by 0.5 year, coupled with an earlier growth spurt. This paper also provides the first growth standards for Chinese from birth to 18 years, with percentile charts on both standing height and sexual maturation in boys and girls.

DATASET DESCRIPTION:

This dataset contains only the height (inches) and weights (pounds) of 25,000 different humans of 18 years of age. This dataset can be used to build a model that can predict the heights or weights of a human. Childhood obesity is an emerging problem in Asia. Sequential monitoring of the growth of an individual can detect a change in body fatness, provided there are ethnically appropriate growth references.

PROBLEM STATEMENT:

Build a predictive model for determining height or weight of a person, implement a linear regression model that will be used for predicting height or weight

LIBRARIES:

We have used some of the standard libraries for the respective dataset for regression analysis and the libraries are as shown

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

TASKS

- DATA VISUALIZATION
- DATA MODELLING
- TESTING
- Results and conclusion

DATA VISUALIZATION:

The graphical representation of information and data is known as data visualisation. Data visualisation tools make it easy to see and understand trends, outliers, and patterns in data by using visual elements like charts, graphs, and maps. There are many types of data visualization. The most common are scatter plots, line graphs, pie charts, bar charts, heat maps, area charts, choropleth maps and histograms. This procedure examines the dataset structure, looks for potential data problems, and provides a clear understanding of the data. The data gathered in this section can be applied to the modelling phase. Although UCI provides the dataset ready for analysis, it's a good idea to double-

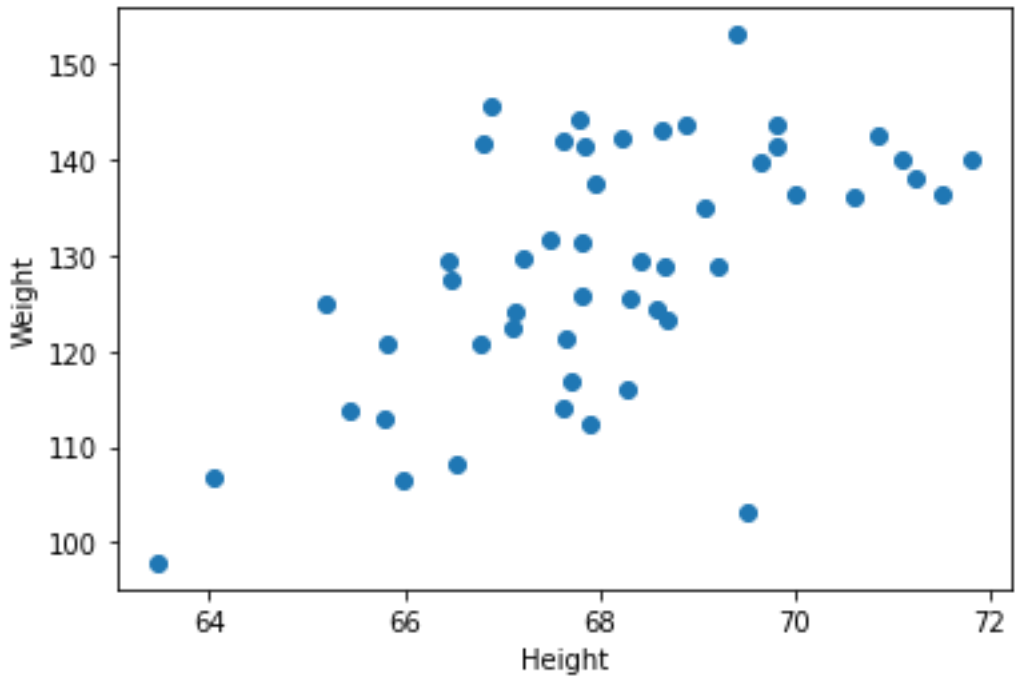
check it for any potential problems. There are five sub-processes to follow when preparing data. Data selection, cleansing, construction, integration, and formatting are the steps in the process. To put it another way, all of these steps encompass all of the activities that must be completed in order to construct the final data set.

SOCR-Height Weight Dataset: The Height-Weight Dataset in Rows and Columns is represented as

```
df = pd.read_csv('/content/SOCR-HeightWeight.csv')
df
```

	Index	Height(Inches)	Weight(Pounds)
0	1	65.78331	112.9925
1	2	71.51521	136.4873
2	3	69.39874	153.0269
3	4	68.21660	142.3354
4	5	67.78781	144.2971
...	...	...	...
24995	24996	69.50215	118.0312
24996	24997	64.54826	120.1932
24997	24998	64.69855	118.2655
24998	24999	67.52918	132.2682

VISUALIZED DATA SET WITH HEIGHT IN X-AXIS AND WEIGHT IN Y-AXIS USING SCATTER PLOT:



Here in data visualisation we have mainly used scatter plot ,it is a type of a plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data. The main application why we have chosen the particular plot is, scatter plot useful for identifying other patterns in data. Firstly, we divide data points into groups based on how closely sets of points clusters together so that it can also show if there are any unexpected gaps in the data and if there are any outlier points. This can be useful if we want to segment the data into different parts, like in the development of user personas. In order to create a scatter plot, we need to select two columns from a data table, one for each dimension of the plot. Each row of the table will become a single dot in the plot with position according to the column values. The code for Scatter plot is as shown:

```
plt.scatter(x[:50],y[:50])
plt.xlabel('Height')
plt.ylabel('Weight')
plt.show()
```

DATA MODELLING:

The practise of examining data objects and their relationships with other things is known as data modelling. It's utilised to investigate the data requirements for various business activities. The data models are constructed to store the information in a database. Instead, then focusing on what processes we must conduct, the Data Model focuses on what data is required and how we must organise it.

For data modelling we have used regression methods.

Regression is a technique for investigating the relationship between independent variables or features and a dependent variable or outcome. It's used as a method for predictive modelling in machine learning, in which an algorithm is used to predict continuous outcomes. Solving regression problems is one of the most common applications for machine learning models, especially in supervised machine learning. Algorithms are trained to understand the relationship between independent variables and an outcome or dependent variable. The model can then be leveraged to predict the outcome of new and unseen input data, or to fill a gap in missing data. Regression analysis is an integral part of any forecasting or predictive model, so is a common method found in machine learning powered predictive analytics. Alongside classification, regression is a common use for supervised machine learning models. This approach to training models required labelled input and output training data. Machine learning regression models need to understand the relationship between features and outcome variables, so accurately labelled training data is vital. Regression is a key element of predictive modelling, so can be found within many different applications of machine learning. Whether powering financial forecasting or predicting healthcare trends, regression analysis can bring organisations key insight for decision-making.

Linear Regression:

Linear regression is a method where a straight line is used to determine the relationship between input and output values. Predictions are made as a combination of the input values to predict the output value. Each input point (x) is weighted using a coefficient (b), and the goal of the algorithm is to locate a set of coefficients that results in good predictions (y)

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + ...$$

Testing and Training:

Firstly, we will take the dimension of X and Y component which is Height and Weight respectively as shown

```
x.ndim, y.ndim
```

```
(2, 1)
```

For modelling purpose, we have to take x_train, x_test,y_train, y_test for predicting the values and find the predicted model over actual.

```
x_train,x_test,y_train,y_test = train_test_split(x,y,random_state = 0, test_size = 0
```

```
model = LinearRegression()
model.fit(x_train, y_train)
y_pred = model.predict(x_test)
```

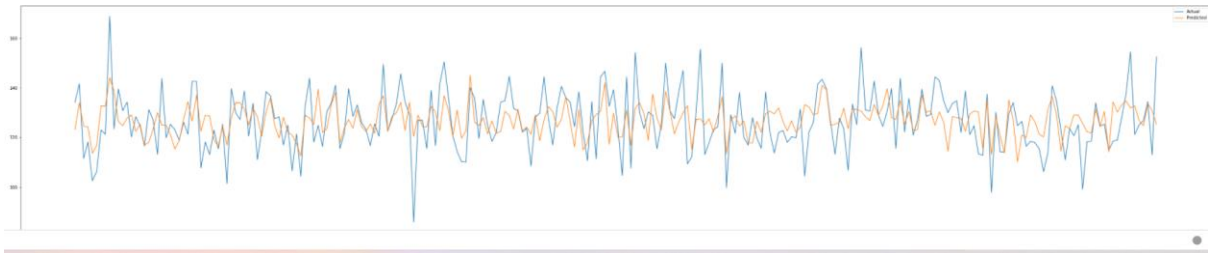
```
from sklearn.metrics import r2_score
r2_score(y_test,y_pred)
```

```
0.26003111920352195
```

Results and Accuracy:

We can successfully predict the weights of a particular person considering height as a reference.

```
df1.plot(figsize=(50,10)) #pandas plotting
plt.show()
```

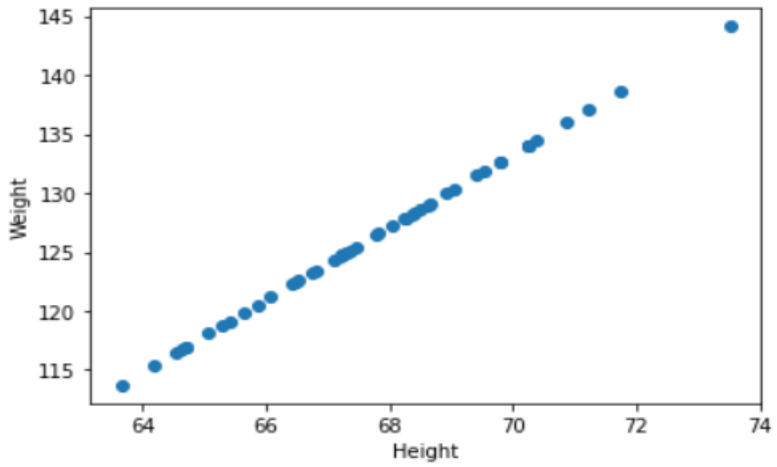


Before starting the predictions, the report makes a summary of model evaluation, explaining the most common metrics used in categorical problems in machine learning.

In data preparation, the training and testing sets are created, and they will be used during the model building.

In data exploration and visualization, we look for features that may provide good prediction results. The best predictors have low distribution overlapping area and low correlation among them.
Modelling starts explaining very simple models and gradually moves to more complex ones. There's a brief explanation on some of the models. used in this report

```
plt.scatter(x_test[:50],y_pred[:50])
plt.xlabel('Height')
plt.ylabel('Weight')
plt.show()
```

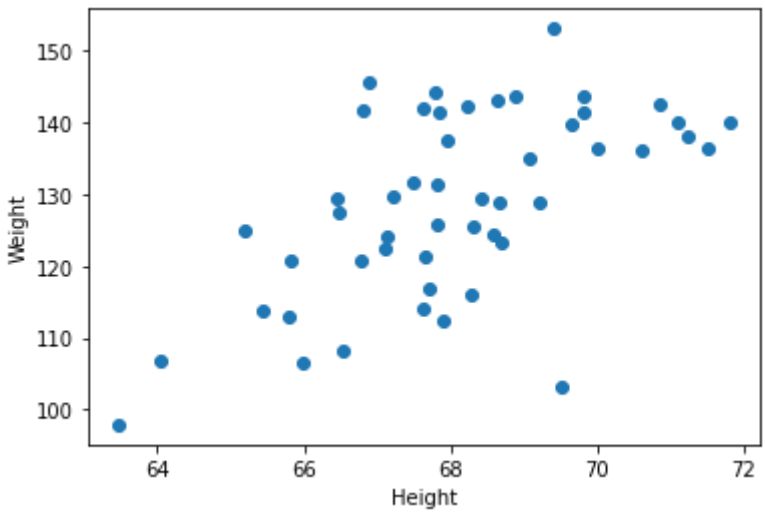


CODE

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
```

+ Code + Text

```
df = pd.read_csv('/content/SOCR-HeightWeight.csv')

x = df[['Height(Inches)']].values

y = df['Weight(Pounds)'].values

plt.scatter(x[:50],y[:50])
plt.xlabel('Height')
plt.ylabel('Weight')
plt.show()
```



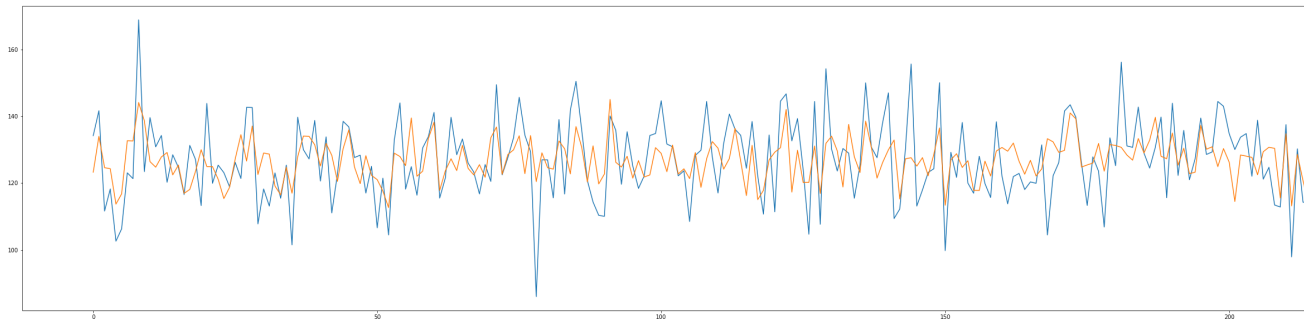

```
x_train,x_test,y_train,y_test = train_test_split(x,y,random_sta

model = LinearRegression()
model.fit(x_train, y_train)
y_pred = model.predict(x_test)
```

```
from sklearn.metrics import r2_score
print(r2_score(y_test,y_pred))
```

```
0.26003111920352195
```

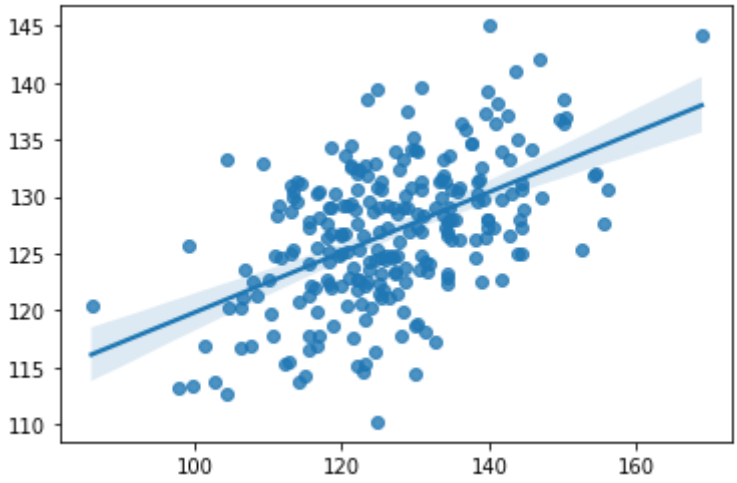
```
df1 = pd.DataFrame({'Actual' :y_test, 'Predicted' : y_pred})
```

```
df1.plot(figsize=(50,10)) #pandas plotting
plt.show()
```



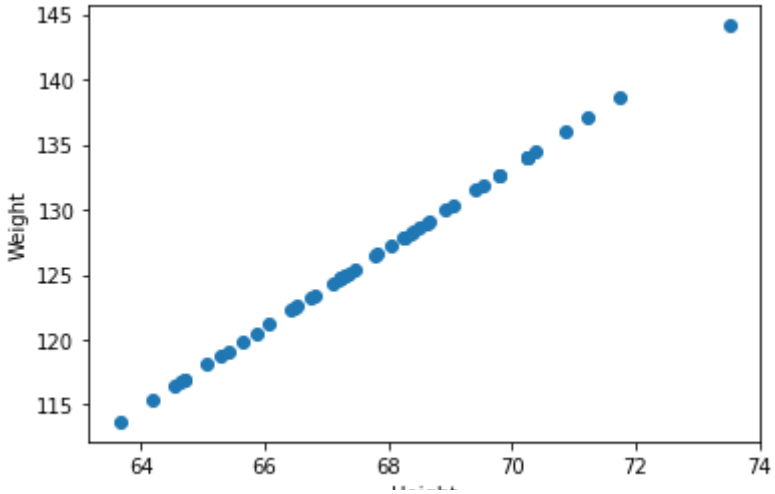
```
sns.regplot(y_test, y_pred)
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning:
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7fd075cae750>
```




```
plt.scatter(x_test[:50],y_pred[:50])
plt.xlabel('Height')
```

```
plt.ylabel('Weight')
plt.show()
```




```
plt.scatter(x_test[:50],y_test[:50])
plt.xlabel('Height')
plt.ylabel('Weight')
plt.show()
```