Big Data Assignment

Jagmeet Singh Sidhu **1102275**

Data:

- The dataset contains the body mass index of people, **'Data.csv'**
- The two columns are '**Gender', 'Index'**, these are basically the gender and their body mass index of the population and both these columns are categorical data
- The other two columns are **'Height', 'Weight'**, these are the continuous data that contains height and weight of the respective people.
- There are total 500 entries
- Source -> https://www.kaggle.com/yersever/500-person-gender-height-weight-bodymassindex/version/2#

Methods:

- I have used three methods for measuring the distance.
- **Euclidean distance -** The Euclidean distance between two points in either the plane or 3-dimensional space measures the length of a segment connecting the two nodes representing distance between two points.

$$= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

- **Manhattan distance -** The distance between two points measured along axes at right angles. In a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2), it is $|x_1 - x_2| + |y_1 - y_2|$

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

.

- **Chi-Square distance** - The chi-square distance has the property of distributional equivalence, meaning that it ensures that the distances between rows and columns are invariant when two columns (or two rows) with identical profiles are aggregated.

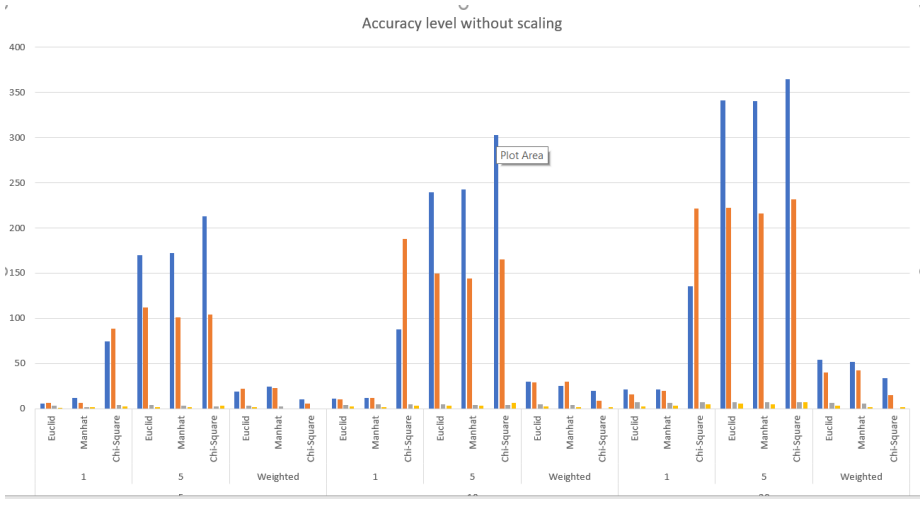
$$d(\mathbf{x}_u, \mathbf{x}_v) = \sum_{n=1}^{N} \frac{[x_u(n) - x_v(n)]^2}{x_u(n) + x_v(n)}.$$

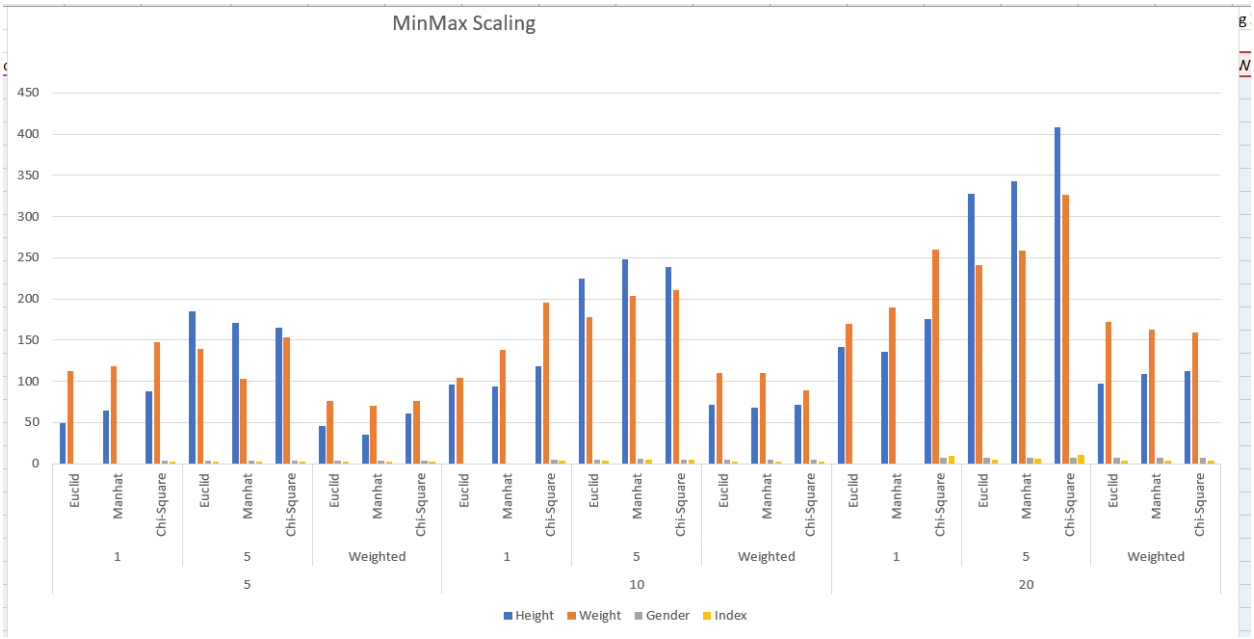
- **Minmax Scaler -** Transforms features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one.
- **Z Score Scaler -** Standardize features by removing the mean and scaling to unit variance. The standard score of a sample x is calculated as, z = (x - u) / s where u is the mean of the training samples or zero if with mean=False, and s is the standard deviation of the training samples or one if with_std=False.
- **Accuracy:** For accuracy I am finding RMSE value which can be founded by the following formula

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Tools:

- Python 3
- Jupyter Notebook
- Anaconda Navigator
- Libraries- Pandas (import pandas as pd)
  NumPy (import pandas as pd)
  Sklearn (import sklearn)
  Random (import random)
  Math (import math)




MinMax Scaling
450
400
350
300
250
200
150
100
50
0
Euclid
Manhat
Chi-Square
1
5
Weighted
5
10
20
Height
Weight
Gender
Index


Z-score Scaling
Plot Area
600
500
400
300
200
100
0
Euclid
Manhat
Chi-Square
1
5
Weighted
5
10
20
Height
Weight
Gender
Index