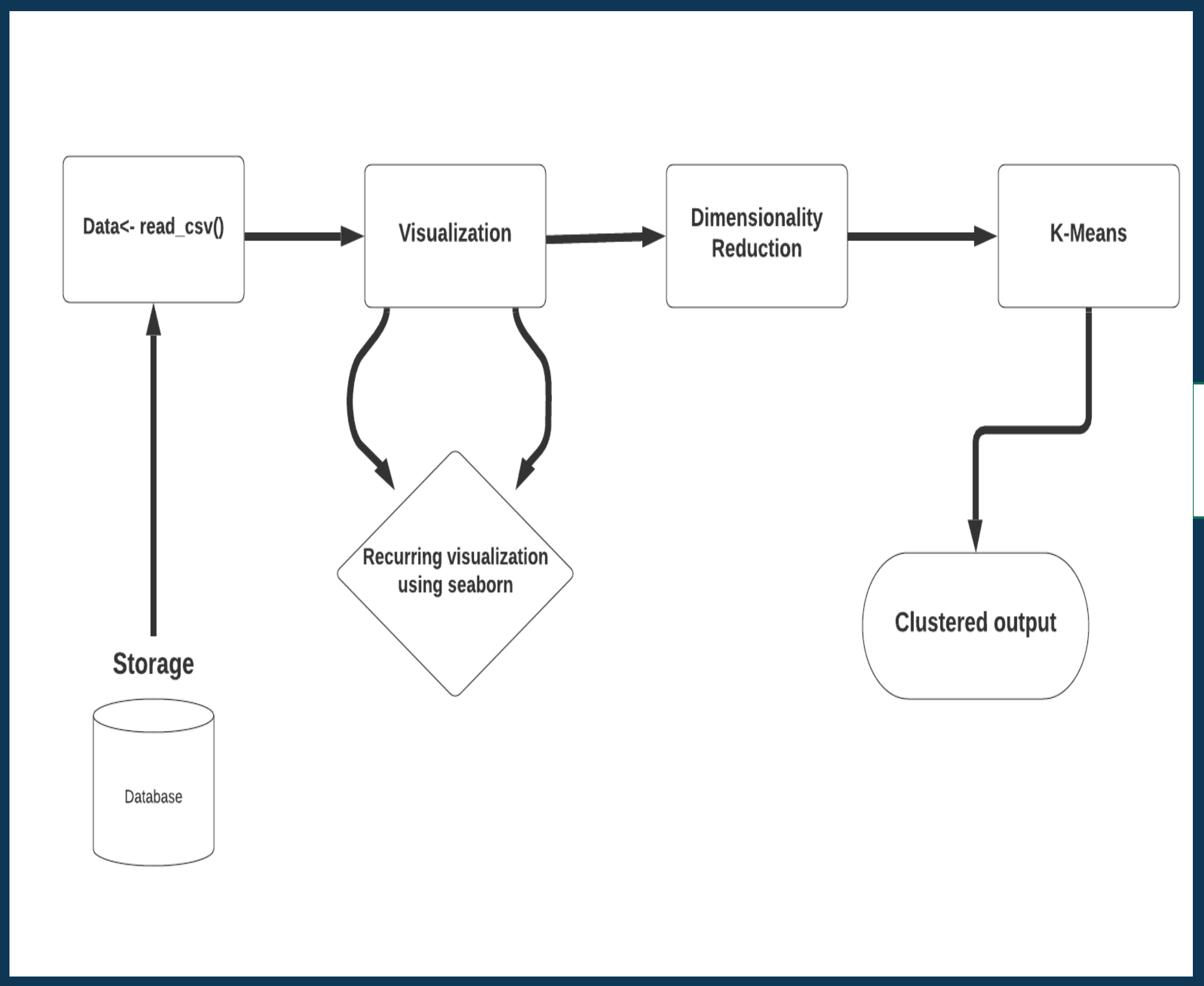
ANALYSIS AND PROCESSING ALL INDIA WEATHER DATA USING AN UNSUPERVISED LEARNING ALGORITHM TO IDENTIFY MINIMUM TEMPERATURE CLUSTERS

ASHA GEORGE
SUPERVISOR: Dr. TODD COCHRANE

Nelson Marlborough Institute of Technology
Te Whare Wānanga o Te Tau Ihu o Te Waka a Maui

System Architecture



Data<- read_csv()

Visualization

Dimensionality Reduction

K-Means

Recurring visualization using seaborn

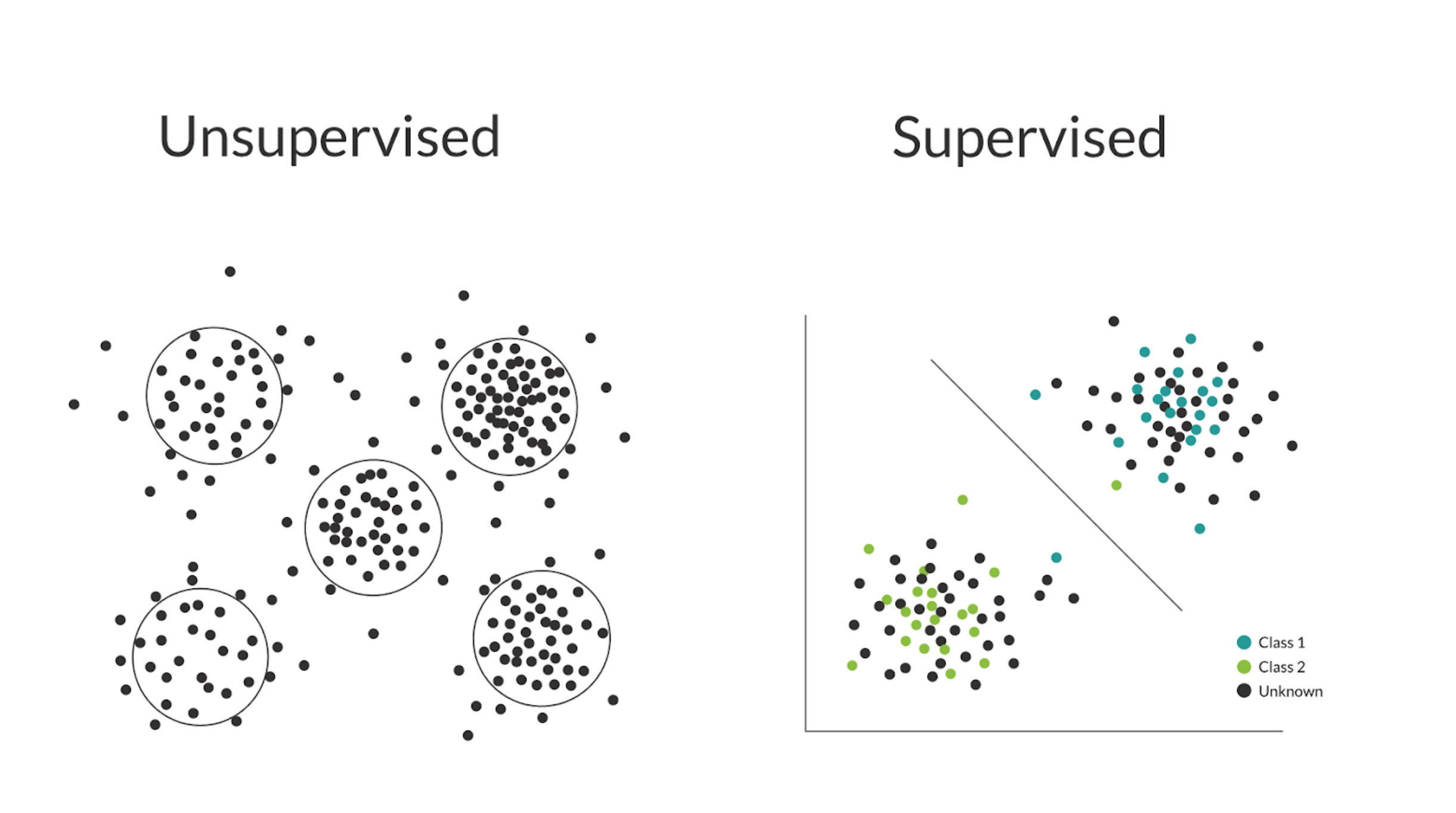
Clustered output

Storage

Database

UNSUPERVISED LEARNING

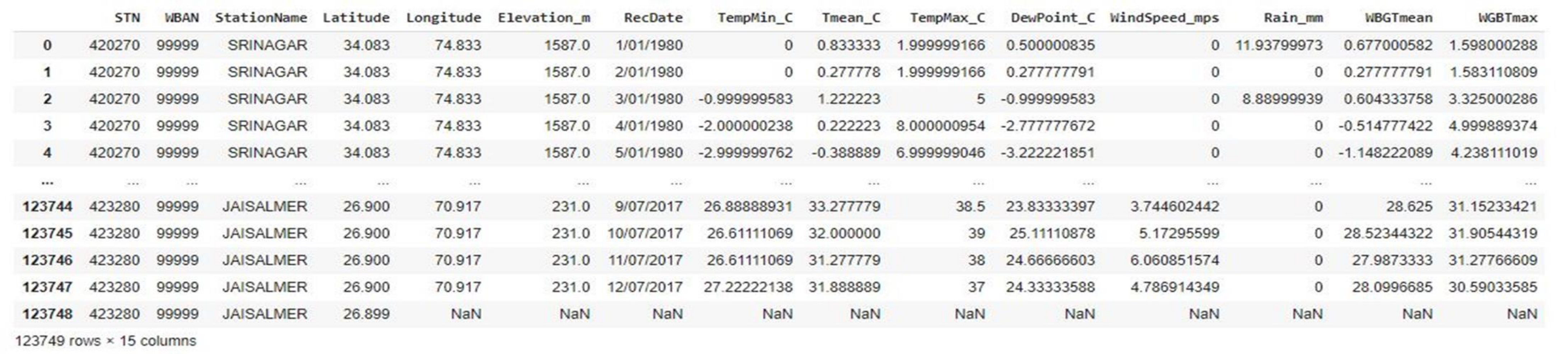
An unsupervised algorithms can work with unlabeled data. In other words, the system itself will be able to improve its intelligence with respect to an inbuilt hidden pattern.



Clustering is an unsupervised learning method. With a variety of clustering algorithms present, the k-means algorithm is a notable and easy to understand clustering algorithm.

Accumulation of the appropriate dataset

The dataset selected for clustering was WeatherIndiaDaily1980-2019 Dataset. Which is having almost all-weather data like, minimum and maximum temperature, quantity of rain, latitude, longitude.

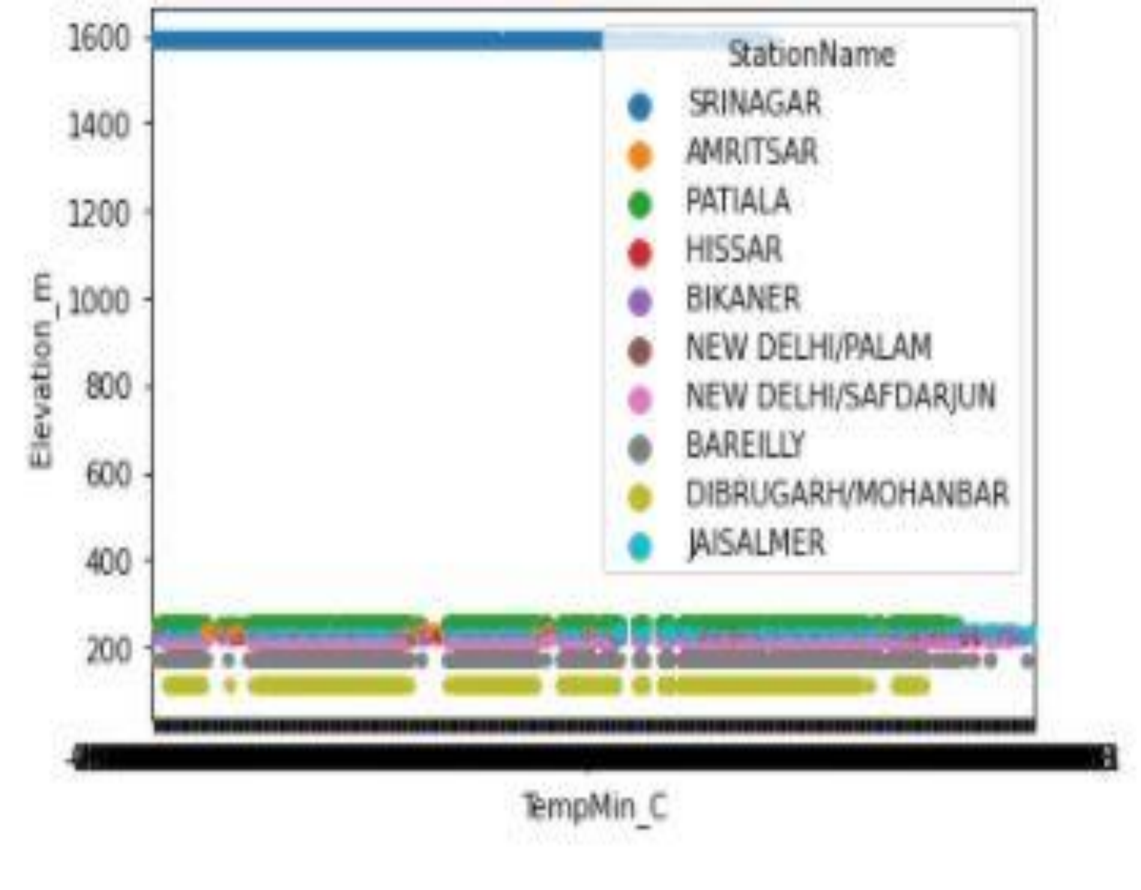


	STN	WBAN	StationName	Latitude	Longitude	Elevation_m	RecDate	TempMin_C	Tmean_C	TempMax_C	DewPoint_C	WindSpeed_mps	Rain_mm	WBGTmean	WGBTmax
0	420270	99999	SRINAGAR	34.083	74.833	1587.0	1/01/1980	0	0.833333	1.999999166	0.500000835	0	11.93799973	0.677000582	1.598000288
1	420270	99999	SRINAGAR	34.083	74.833	1587.0	2/01/1980	0	0.277778	1.999999166	0.277777791	0	0	0.277777791	1.583110809
2	420270	99999	SRINAGAR	34.083	74.833	1587.0	3/01/1980	-0.999999583	1.222223	5	-0.999999583	0	8.88999939	0.604333758	3.325000286
3	420270	99999	SRINAGAR	34.083	74.833	1587.0	4/01/1980	-2.000000238	0.222223	8.000000954	-2.777777672	0	0	-0.514777422	4.999889374
4	420270	99999	SRINAGAR	34.083	74.833	1587.0	5/01/1980	-2.999999762	-0.388889	6.999999046	-3.222221851	0	0	-1.148222089	4.238111019
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
123744	423280	99999	JAISALMER	26.900	70.917	231.0	9/07/2017	26.88888931	33.277779	38.5	23.83333397	3.744602442	0	28.625	31.15233421
123745	423280	99999	JAISALMER	26.900	70.917	231.0	10/07/2017	26.61111069	32.000000	39	25.11110878	5.17295599	0	28.52344322	31.90544319
123746	423280	99999	JAISALMER	26.900	70.917	231.0	11/07/2017	26.61111069	31.277779	38	24.66666603	6.060851574	0	27.9873333	31.27766609
123747	423280	99999	JAISALMER	26.900	70.917	231.0	12/07/2017	27.22222138	31.888889	37	24.33333588	4.786914349	0	28.0996685	30.59033585
123748	423280	99999	JAISALMER	26.899	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

123749 rows × 15 columns

Data Visualization

The data was visualized using Seaborn and Matplotlib to understand the trends within the data from a visual perspective and to have a rough estimate of how balanced the data was.



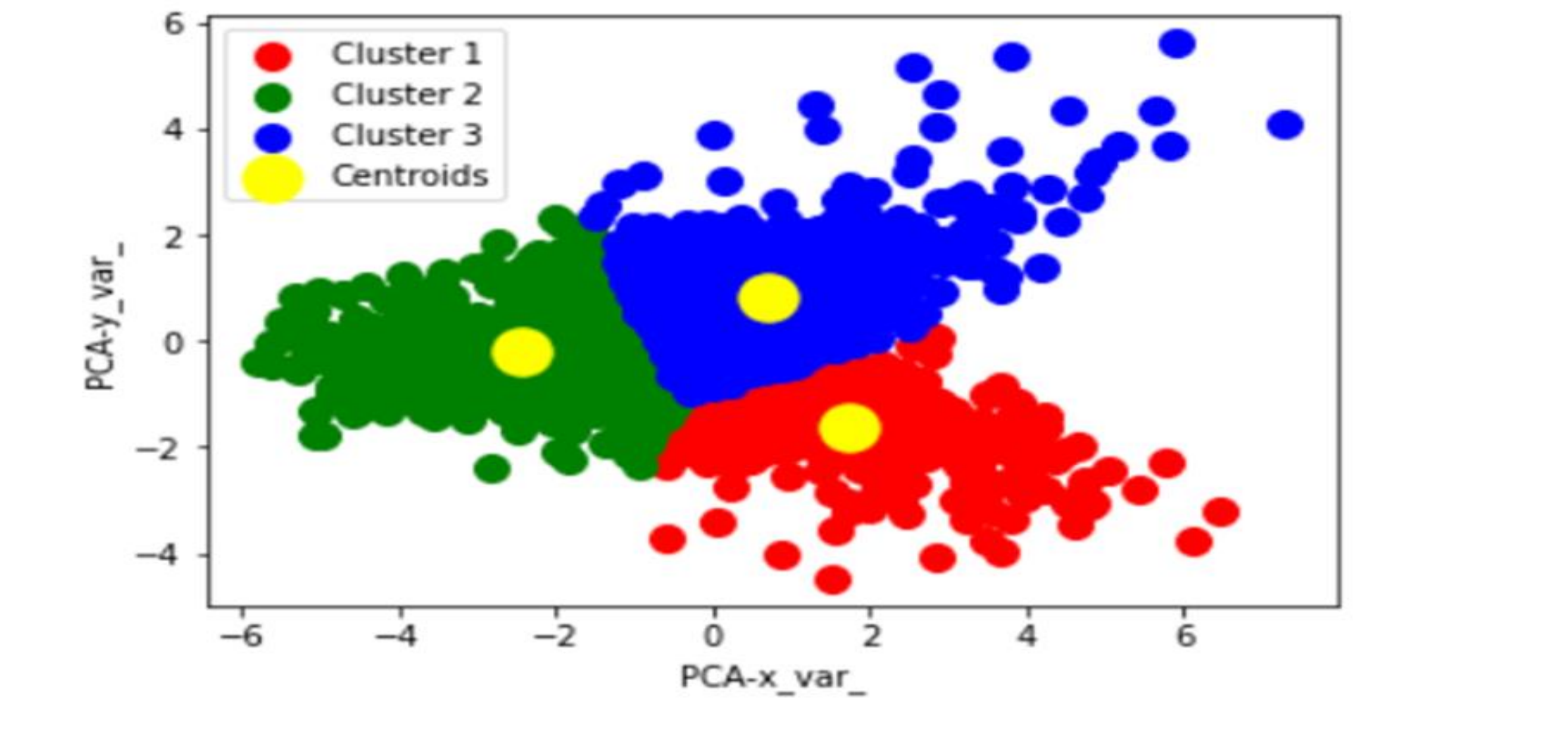
Strip plot

Dimensionality Reduction

Dimensionality Reduction techniques were used to shorten down the overall number of features for faster computation and better visualization purposes.

Clustering using K-Means

The clustering of dataset has done with technique called K-Means. In machine learning K-Means is mainly used for cluster analysis. The dataset clustered into three clusters and plotted as shown in figure.



Asha-George@live.nmit.ac.nz