DR. BR AMBEDKAR NATIONAL INSTITUTE OF TECHNOLOGY, JALANDHAR (144011), PUNJAB

Machine Learning (CSPC-204)

Mini-Project

Submitted by:

Rajan Dhiman (20103117)

Rashmi Kumari (20103121)

Shruti Gupta (20103135)

Branch- CSE (B - G2)

B.Tech – 2nd Year

Submitted to:

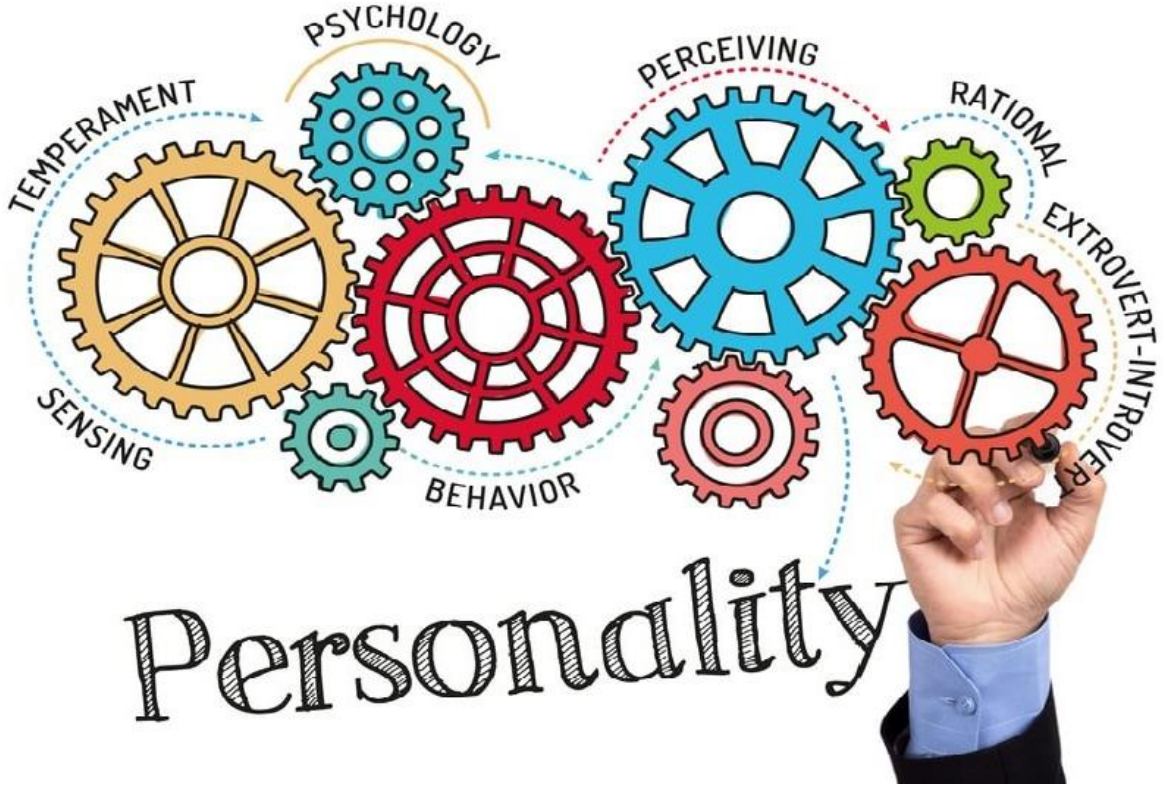
Dr Jagdeep Kaur

Introduction

Personality of a person encircles every aspect of life. It describes the pattern of thinking, feeling and characteristics that predict and describe an individual's behaviour and also influences daily life activities including emotions, preference, motives and health.

Nowadays personality recognition from social networking sites has attracted the attention of researchers for developing automatic personality recognition systems. The core philosophy of such applications is based on the different personality models, like Big Five Factor Personality Model.

Our model predicts the personality type of an individual depending on a specific set of questions for each of the five attributes i.e. Extroversion, Nervousness, Compatibility, Meticulousness and Creativity.



Problem Statement

Predicting personality from online text is a growing trend for researchers. Sufficient work has already been carried out on predicting personality from the input text. However, more work is required to be carried out for the performance improvement of the existing personality recognition system, which in most of the cases arises due to presence of imbalanced classes of personality traits. In the proposed work, a dataset balancing technique, called re-sampling is used for balancing the personality recognition dataset, which may result in improved performance.

Algorithm Description

K-means clustering algorithm

K-means clustering is the most commonly used clustering algorithm. It's a centroid-based algorithm and the simplest unsupervised learning algorithm.

This algorithm tries to minimize the variance of data points within a cluster.

K-means is best used on smaller data sets because it iterates over *all* of the data points. That means it'll take more time to classify data points if there are a large amount of them in the data set.

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ so as to minimize the within-cluster sum of squares (i.e. variance).

Input:

D={t1 , t2 , … tn} (Set of elements)

K - Number of desired clusters

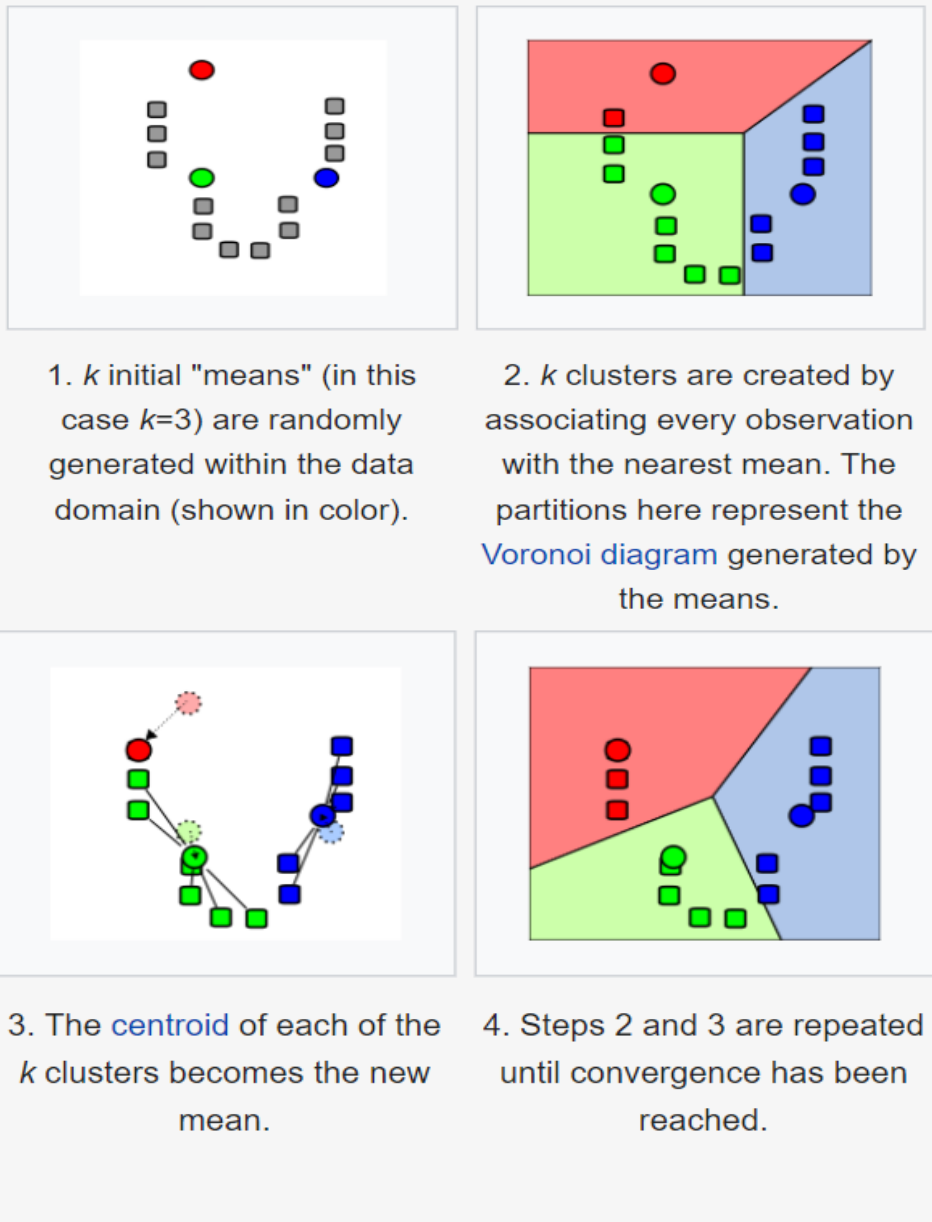
Output:

K – Set of clusters

Pseudo Code of K-Means Algorithm:

1. Specify number of K-clusters to assign
2. Randomly initialize k centroids.
3. **Repeat**
4. **Expectations**: Assign each point to it's closest centroid.
5. **Maximization**: Compute the new centroid (mean) of each cluster.
6. **Until**: the centroid positions do not change.

Demonstration of the standard algorithm



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

The algorithm does not guarantee convergence to the global optimum. The result may depend on the initial clusters. As the algorithm is usually fast, it is common to run it

multiple times with different starting conditions. However, worst-case performance can be slow: in particular certain point sets, even in two dimensions, converge in exponential time, that is $2^{\Omega(n)}$. These point sets do not seem to arise in practice: this is corroborated by the fact that the smoothed running time of *k*-means is polynomial.

We implemented k-Means in our project by dividing the dataset into 10 clusters, each of batch size 100 and maximum iterations as 100. These clusters are having similar personality traits, and are used to separately determine the distribution of different traits on a fixed set of inputs.

Aims and Objective

i. **Aim:**

The aim of this work is to classify the personality traits of a user from the input text by applying unsupervised machine learning technique namely mini-batch K-means algorithm on the BIG FIVE PERSONALITY TEST dataset.

ii. Objectives:

a. Applying machine learning technique namely K-means classifier for personality traits recognition from the input text.
b. Learning to apply normalization on the imbalanced classes of personality traits for improving the performance of proposed system.

Dataset Description

The personality test was constructed with the "Big-Five Factor Markers" from the IPIP.
https://www.kaggle.com/datasets/tunguz/big-five-personality-test

This dataset contains 110 columns and 1,015,342 questionnaire answers collected online by Open Psychometrics.

The following items were presented on one page and each was rated on a five point scale using radio buttons. The order on page was EXT1, AGR1, CSN1, EST1, OPN1, EXT2, etc.

The scale was labelled as:

1 = Disagree, 3 = Neutral, 5 = Agree

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	...	dateload	screenw	screenh	introelapse	testelapse	endelapse	IPC	country	lat_appx_lots_of_
0	4.0	1.0	5.0	2.0	5.0	1.0	5.0	2.0	4.0	1.0	...	2016-03-03 02:01:01	768.0	1024.0	9.0	234.0	6	1	GB	51.54
1	3.0	5.0	3.0	4.0	3.0	3.0	2.0	5.0	1.0	5.0	...	2016-03-03 02:01:20	1360.0	768.0	12.0	179.0	11	1	MY	3.16
2	2.0	3.0	4.0	4.0	3.0	2.0	1.0	3.0	2.0	5.0	...	2016-03-03 02:01:56	1366.0	768.0	3.0	186.0	7	1	GB	54.91
3	2.0	2.0	2.0	3.0	4.0	2.0	2.0	4.0	1.0	4.0	...	2016-03-03 02:02:02	1920.0	1200.0	186.0	219.0	7	1	GB	51.
4	3.0	3.0	3.0	3.0	5.0	3.0	3.0	5.0	3.0	4.0	...	2016-03-03 02:02:57	1366.0	768.0	8.0	315.0	17	2	KE	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	

EXT – Extroversion

EST – nervousness

AGR- compatibility

CSN- Meticulousness

OPN – Creativity

In the dataset, Ext1…EXT10 would help use determine how much extroverted a person is by awarding scores or labels to these statements. Similarly, other traits are judged on the same basis.

Example-

EXT1 I am the life of the party.

EST10 I often feel blue.

AGR9 I feel others' emotions.

CSN6 I often forget to put things back in their proper place.

OPN5 I have excellent ideas.

The dataset contains the score obtained by every individual corresponding to each of the above mentioned attributes.

Preprocessing and Feature Selection

Applied normalization on the dataset using the formula given below.

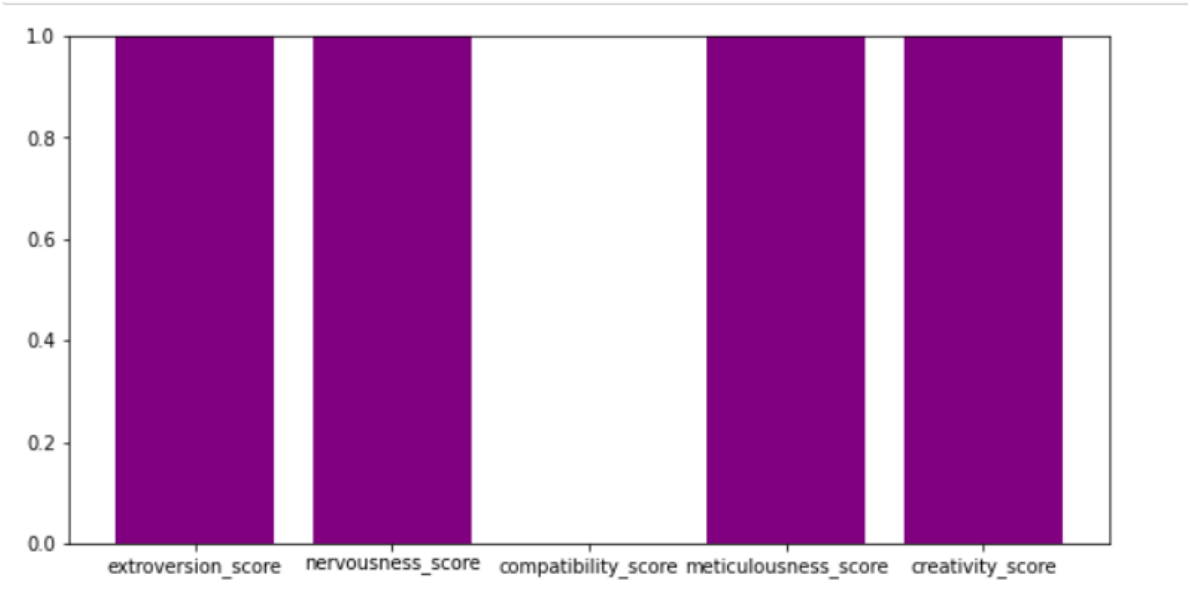
$$X_{normalized} = \frac{(x - x_{minimum})}{(x_{minimum} - x_{minimum})}$$

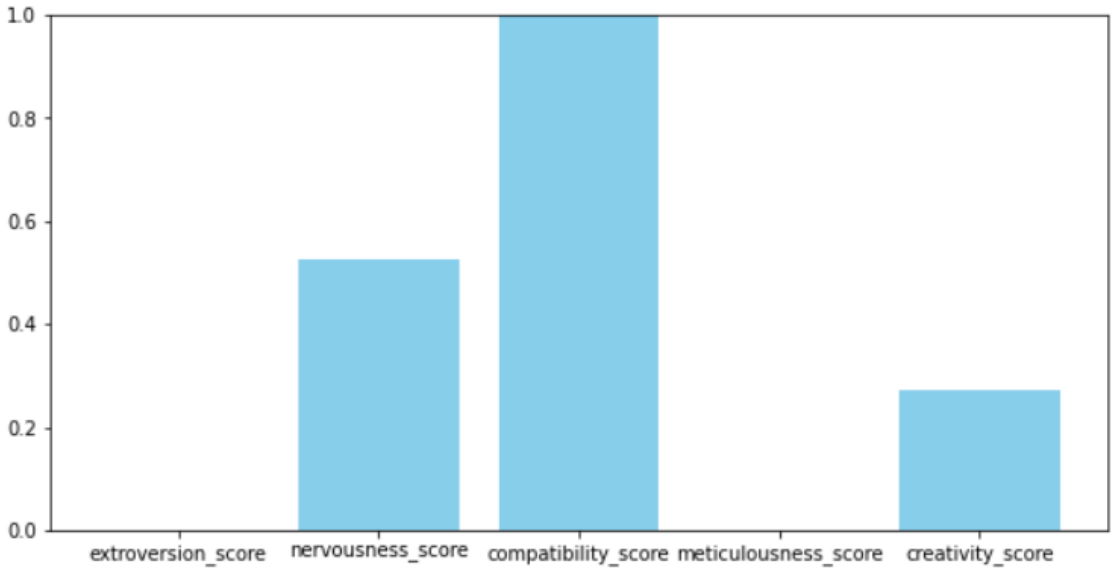
Code snippet-

```
In [34]:
all_extroversion_normalized = (all_extroversion-min(all_extroversion))/(max(all_extroversion)-min(all_extroversion))
all_nervousness_normalized = (all_nervousness-min(all_nervousness))/(max(all_nervousness)-min(all_nervousness))
all_compatibility_normalized = (all_compatibility-min(all_compatibility))/(max(all_compatibility)-min(all_compatibility))
all_meticulousness_normalized = (all_meticulousness-min(all_meticulousness))/(max(all_meticulousness)-min(all_meticulousness))
all_creativity_normalized = (all_creativity-min(all_creativity))/(max(all_creativity)-min(all_creativity))

In [35]:
all_extroversion_normalized

Out[35]: array([1.        , 0.        , 0.58292703, 0.84414155, 0.36153014,
       0.63839681, 0.96798501, 0.27017414, 0.21386007, 0.51772601])
```

Result

Our model represents the overall personality trait of a cluster





These two graphs represent the distribution of score of personality traits of two different clusters as predicted by our model.

In cluster A, majority of people have a low compatibility score as compared to cluster B, where the people have a higher compatibility score. This indicates that these people are not so agreeable and hard to convince.

In the same manner, the meticulousness score of second cluster is lower than that of first, which signifies that these people are hardworking than others.

Our Team

- **Rajan Dhiman** (20103117) – Algorithm explaining and plotting graphs
- **Rashmi Kumari** (20103121) – Choosing the Dataset and pre-processing
- **Shruti Gupta** (20103135) – Project ideation and algorithm selection

References

- (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, No. 3, 2020s
- wikipedia
- javapoint
- simplelearn