CSE 573: Semantic Web Mining

Project Demo

Group 6
Project 21: Personality Classification with Social Media

Group Members:

1. Kampally Sreshta Chowdary, 1224031900
2. Kunj Patel, 1213184152
3. Phani Rohitha Kaza, 1219915970
4. Prakruti Singh Thakur, 1222301340
5. Sai Ajitesh Krovvidi, 1219320388
6. Ankit Sharma, 1219472813

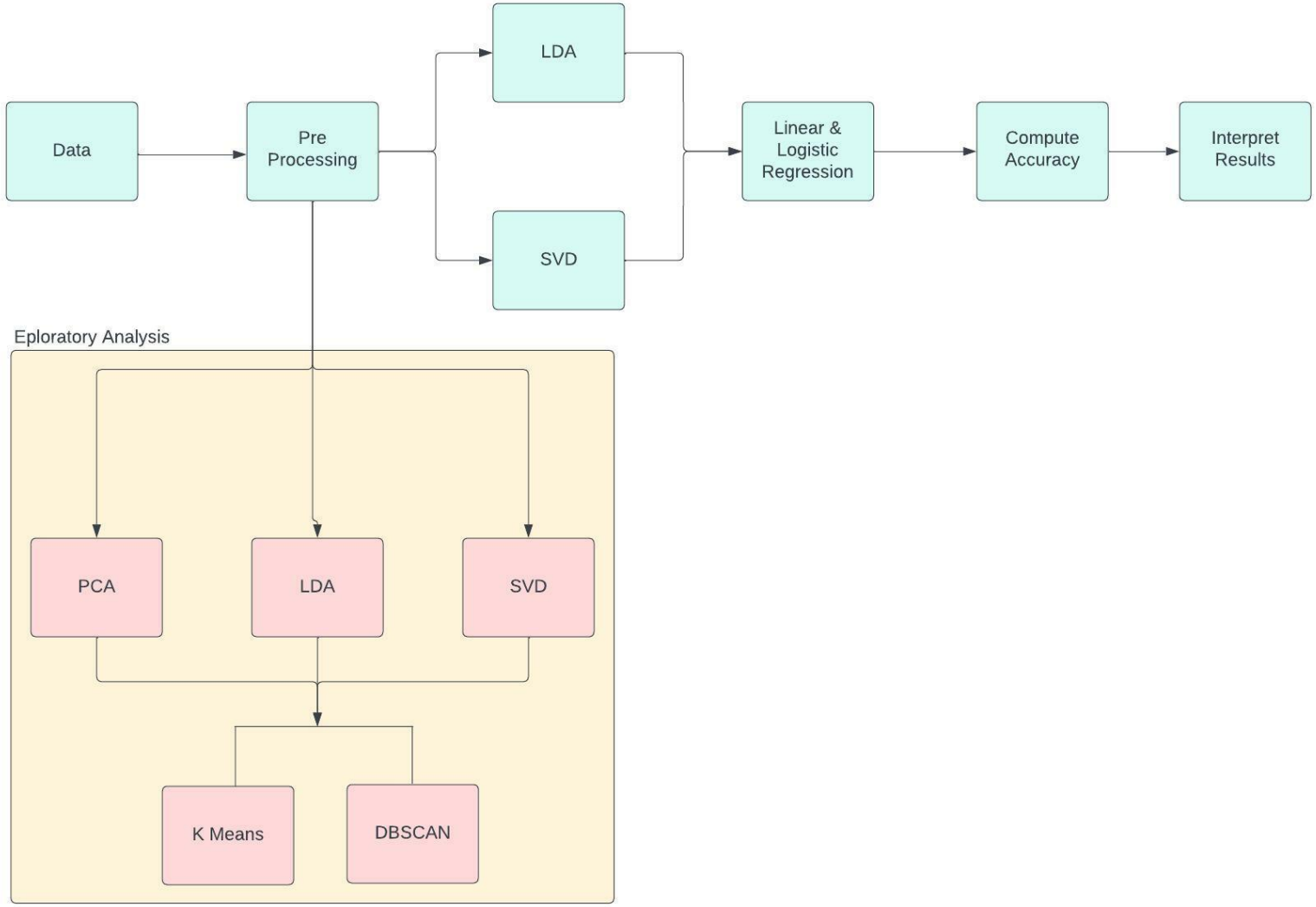
Problem Definition

- With the increasing usage of social media in recent times, it has been possible to get hold of a large amount of data. People interact on social media in many ways. Some of the most popular ways are through liking someone's posts or commenting on them.

- In the Q4 of 2021, the number of daily active users on Facebook reached 1.93 billion. With such a high volume of data availability, it is becoming more and more critical to classify an individual's personality traits to better serve their demands.

- At the same time, it could also translate to more revenue for the company by running targeted ads and marketing campaigns. This is also important because some personality types are more likely to engage in certain types of activities.

System Architecture and Algorithms



Datasets and Preprocessing

The dataset we are using has been obtained from the website- https://www.michalkosinski.com/dataminingtutorial

- **user.csv**: This file contains the data of 110,729 Facebook users. It includes the *userids* and their respective gender, age, political inclination, and 5 personality traits.
- **likes.csv**: This file contains 1,580,284 *likeids* and the name of the post that is liked.
- **users-likes.csv** : This file contains 10,612,326 lines of data. This data comprises two columns which are *userid* and the *likeids* of the posts they liked.

Preprocessing:

- We use sparse matrix with user_ids as rows and like_names as columns. The item in the matrix is '1' if the specific user liked that specific post. Else it is '0'.
- We then remove the users that liked less than 50 posts and like_names that have less than 150 likes.
- The final resultant matrix is of the dimensions 19742x8523. That is, the number of users after preprocessing is 19742 and the number of likes is 8523

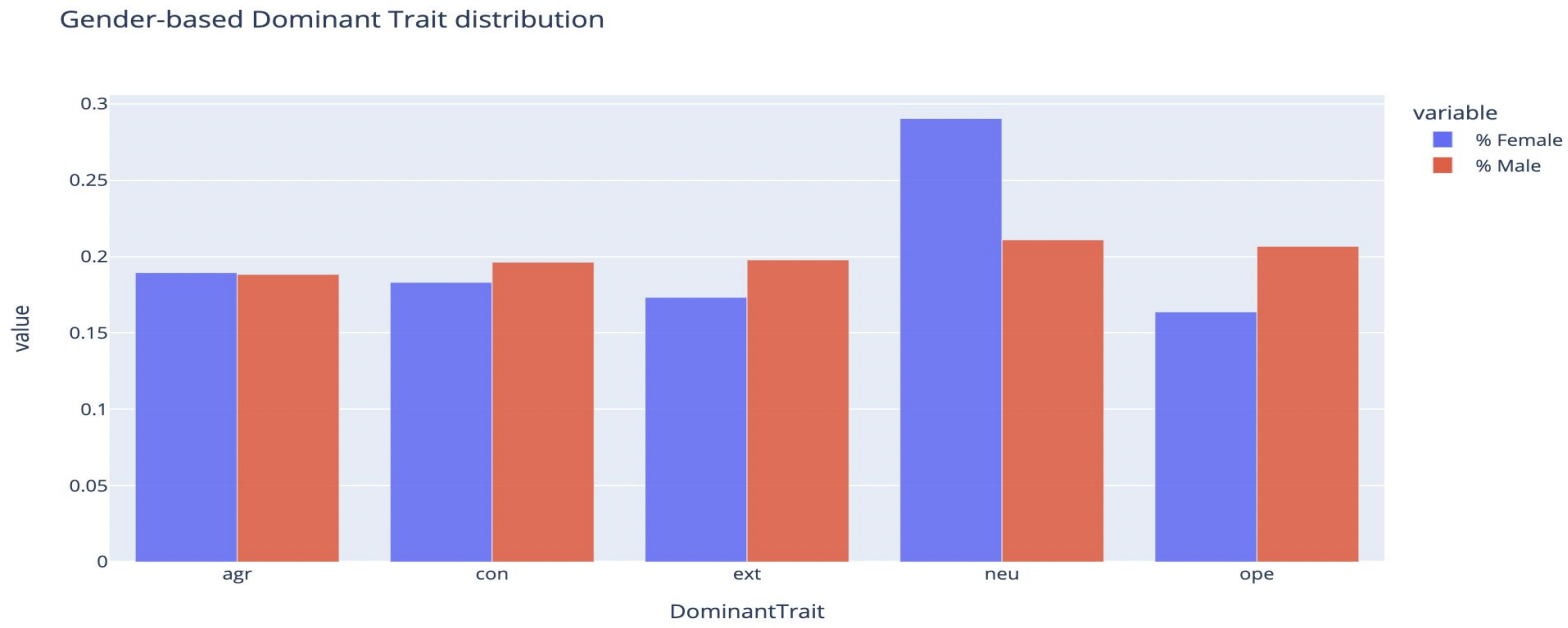
Evaluations

- We compared the accuracies of both the model generated using SVD and LDA.
- For **Openness**, **Conscientiousness**, **Extroversion**, **Neuroticism**, **Agreeableness**, **Gender and Political factor** - SVD had better accuracy.
- For **Age** - LDA gave better results.
- We also compared the accuracies of SVD for different number of dimensions upto 150.

Personality Traits	Highest Accuracy for SVD	Accuracy for LDA(k=150)
Gender	0.955(k=150)	0.87
Age	0.657(k=150)	0.67
Political Factor	0.898(k=80)	0.80
Openness	0.466(k=150)	0.41
Conscientiousness	0.264(k=50)	0.20
Extroversion	0.336(k=150)	0.25
Agreeableness	0.257(k=150)	0.17
Neuroticism	0.317(k=150)	0.25

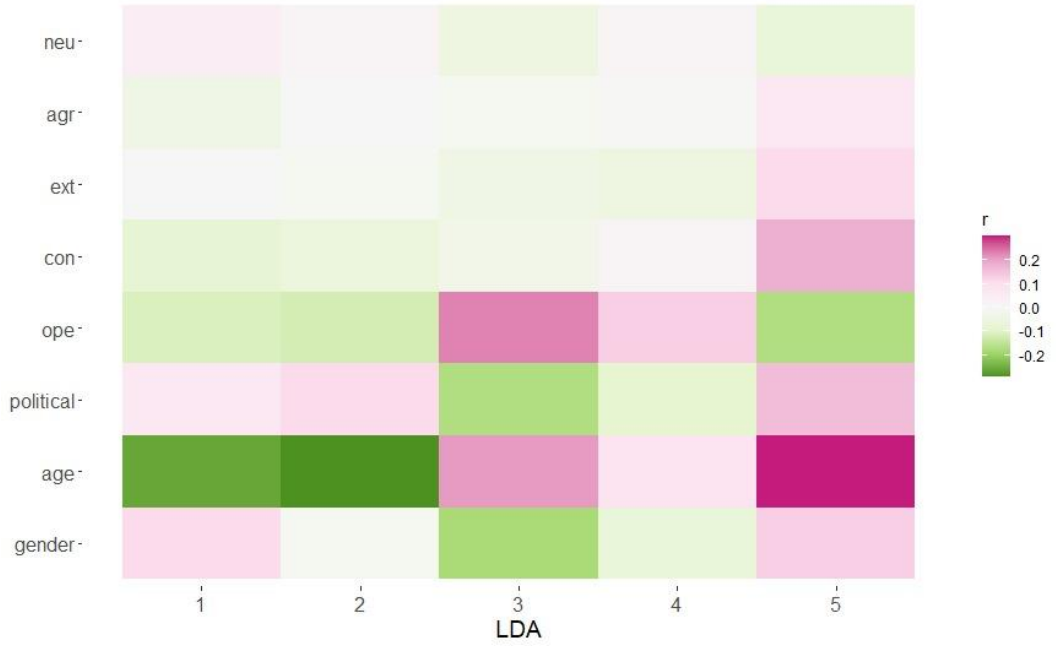
Exploratory Data Analysis

- 110,729 Facebook users - 48773 Males (44%), 61956 Females (56%)
- "Neuroticism" was the most expressed trait among people (especially, in females).
- "Openness" was the least expressed trait (especially, in females)
- Males and Females show the same amount of "Agreeableness"
- Higher percentage of men exhibit "Conscientiousness", "Extraversion" and "Openness" traits than females.

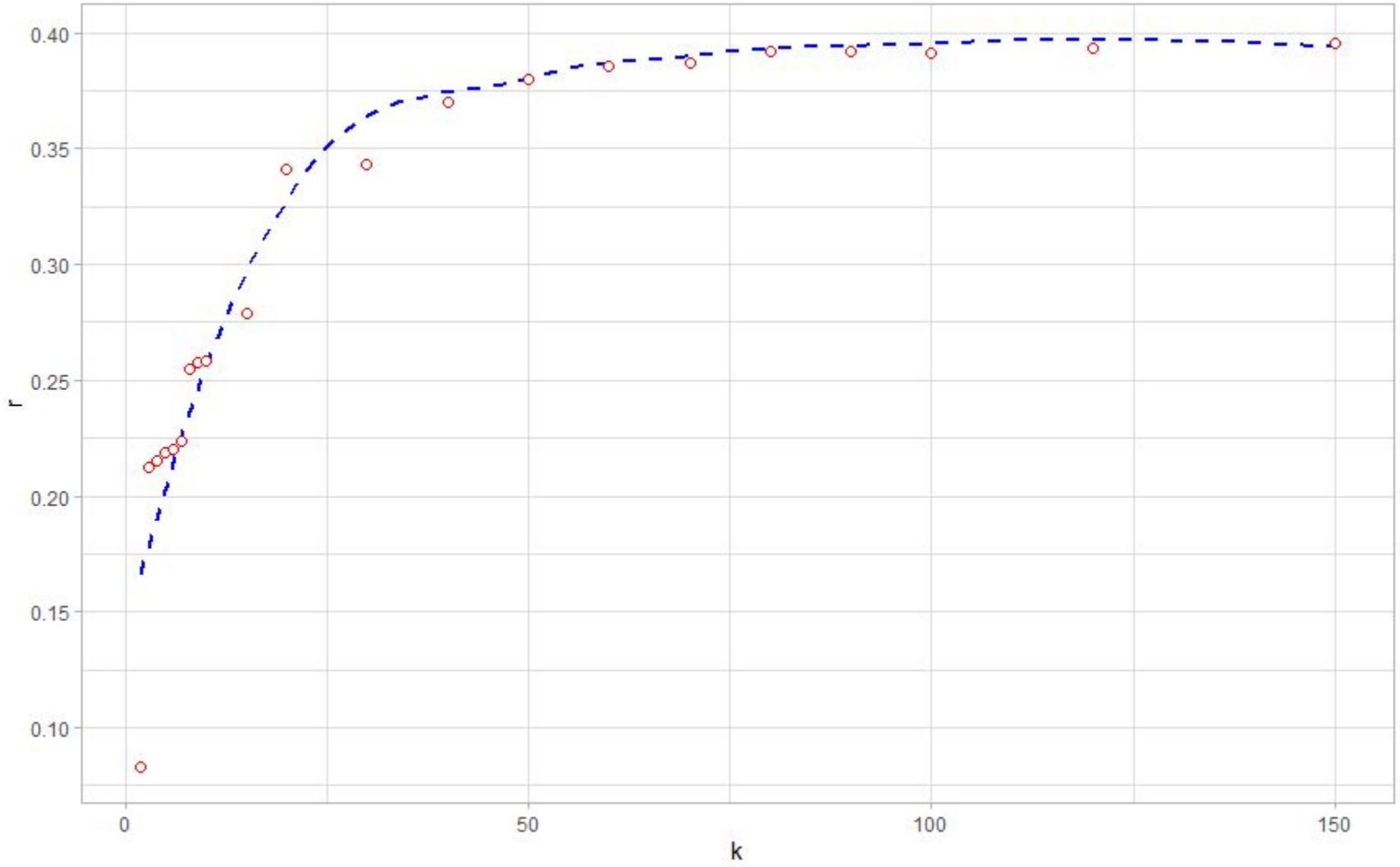


Data Visualization

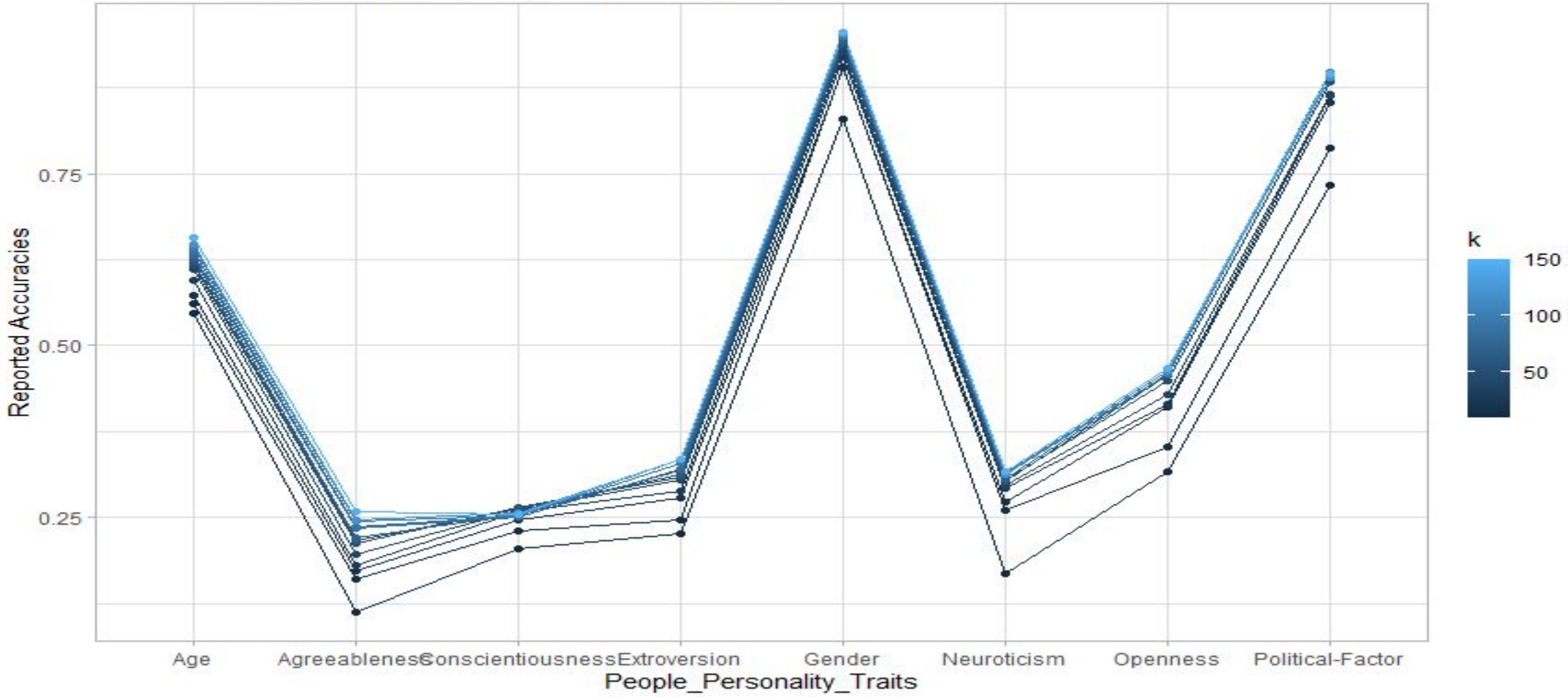
Heat Maps



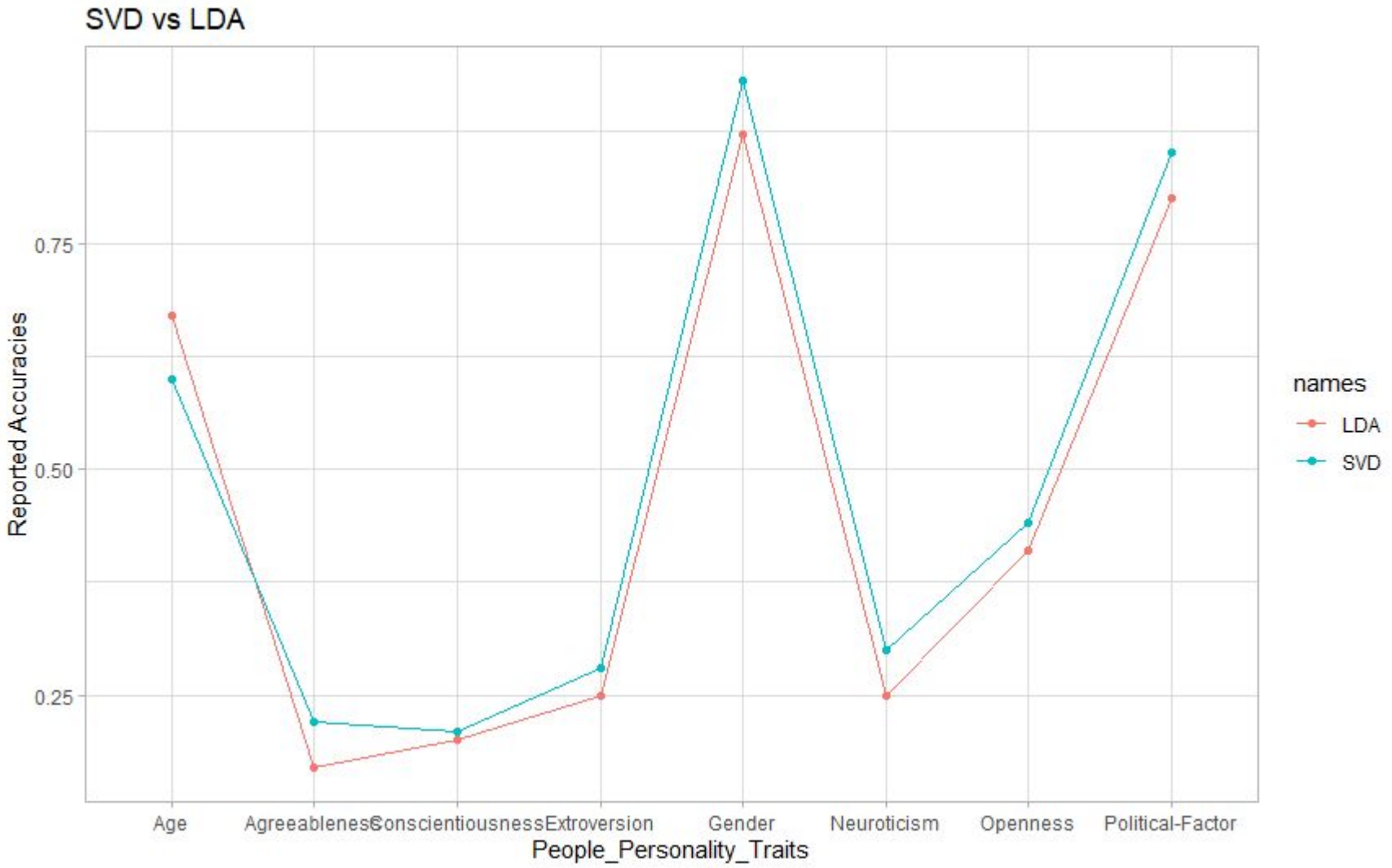
SVD vs Dimensions plot



Prediction plots - Accuracies vs Traits(SVD)

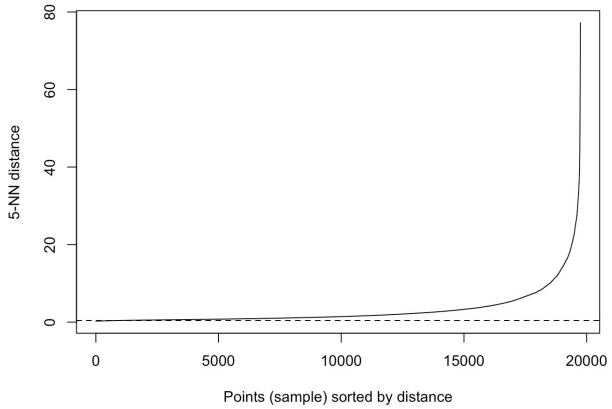


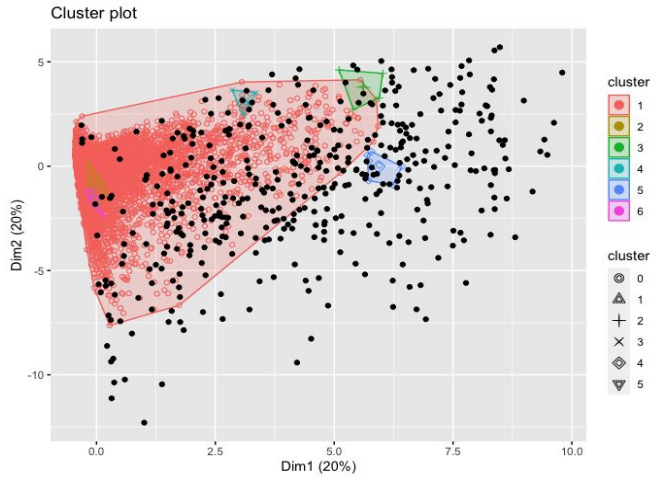
SVD vs LDA plot



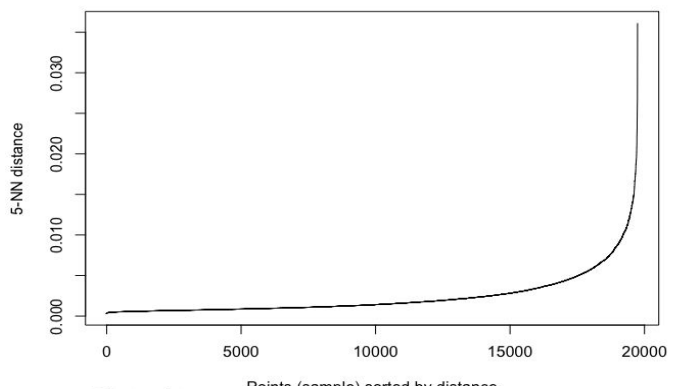
DBSCAN

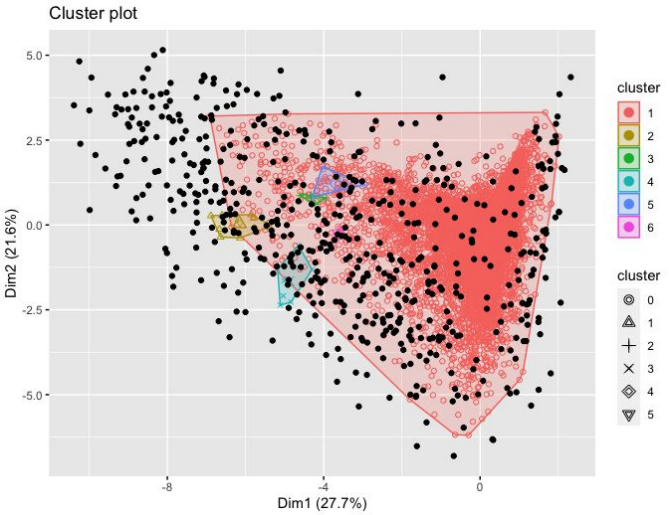
PCA



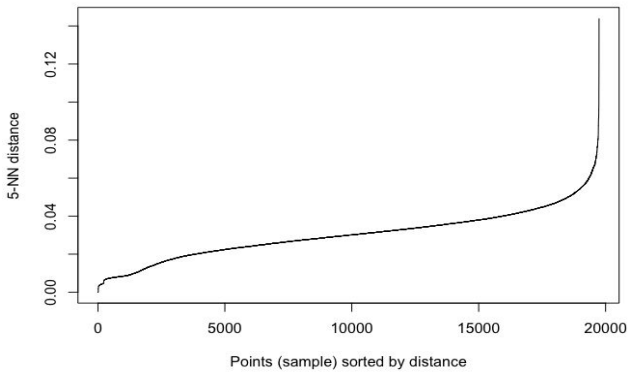


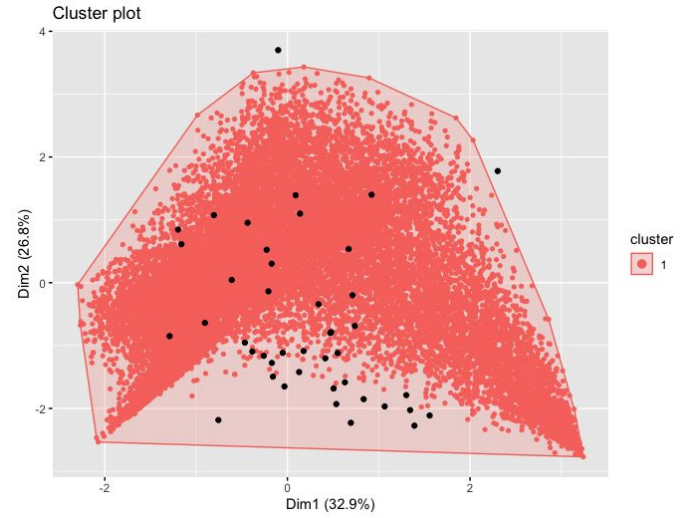
SVD





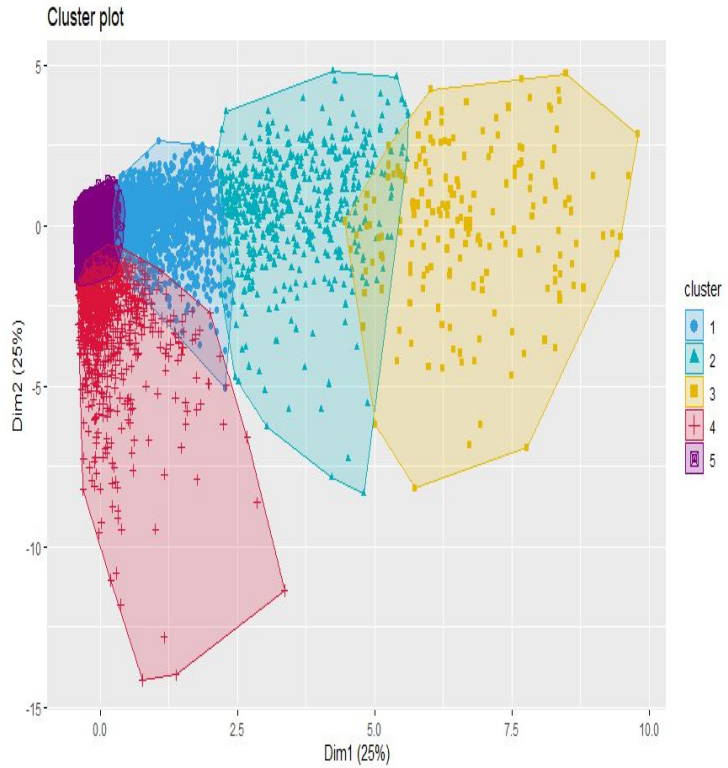
LDA



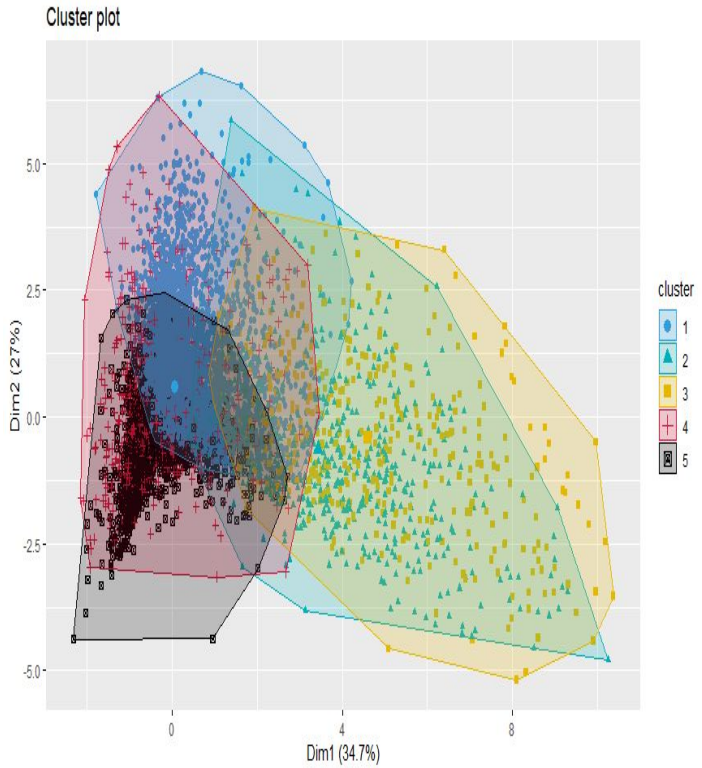


K-means

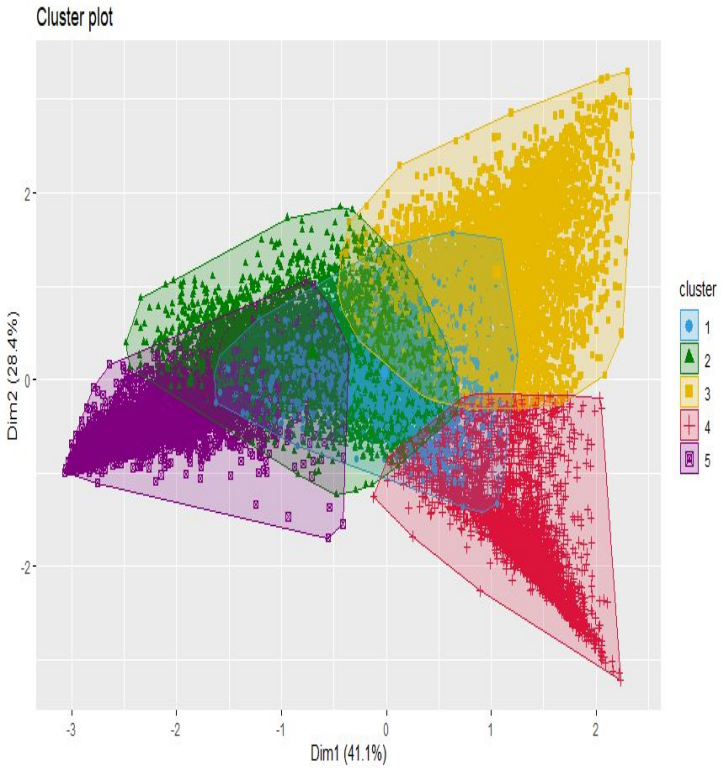
K-means on PCA



K-means on SVD



K-means on LDA



Project Plan: Tasks, Deadlines, Division of Work

Task	Deadline	Assignee
Data Collection & Pre-processing	02-25-2022	Sai Ajitesh Krovvidi, Kunj Patel, Prakruti Singh Thakur
Data Analysis & Visualization	03-09-2022	Phani Rohitha Kaza, Kunj Patel, Prakruti Singh Thakur
Research on Dimensionality Reduction & Clustering Algorithms	03-17-2022	Sai Ajitesh Krovvidi, Sreshta Chowdary Kampally, Ankit Sharma
Perform the Dimensionality Reduction & Clustering Algorithms	03-25-2022	Sai Ajitesh Krovvidi, Sreshta Chowdary Kampally, Kunj Patel
Building the model for predictive Analysis	04-07-2022	Phani Rohitha Kaza, Sreshta Chowdary Kampally, Ankit Sharma
Compare the results of the various Classification models and Choose the appropriate model	04-11-2022	Phani Rohitha Kaza, Sreshta Chowdary Kampally, Prakruti Singh Thakur

References

[1] Michal Kosinski, Yilun Wang, Himabindu Lakkaraju, and Jure Leskovec, "Mining Big Data to Extract Patterns and Predict Real-Life Outcomes", 2016.

[2] Michal Kosinski, https://www.michalkosinski.com/data-mining-tutorial, R code

[3] Dimensionality reduction techniques, 11-dimensionality-reduction-techniques-you-shouldknow-in-2021

[4] Clustering Algorithms, 17 Clustering Algorithms Used In Data Science and Mining

[5] Golbeck, Jennifer, et al. "Predicting personality from twitter." Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011.

[6] Quercia, Daniele, et al. "Our Twitter profiles, our selves: Predicting personality with Twitter."Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third InernationalConference on Social Computing (SocialCom), 2011 IEEE Third International Conference on. IEEE, 2011.

[7] OCEAN traits the-big-five-personality-dimensions

[8] Statista Facebook DAU statistic

[9]https://blog.adioma.com/5-personality-traits-infographic/

Github

https://github.com/KampallySreshtaChowdary/CSE573-Personality-Classification-with-Social-Media

THE END