

Analisis Perbandingan penyebab stress di Negara Maju dan Berkembang berdasarkan aspek sosial menggunakan clustering dan Regresi logistik

OLEH KELOMPOK 3

Agatha Ulina Silalahi
Ivana Joice Chandra
Mastika
Wilsen Vesakha L
Yosua Walfried

230627885
2306174955
2306174974
2306288944
2306175005



Table of contents



01

Pendahuluan

02

Landasan Teori

03

Metodologi

04

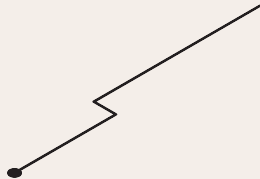
**Hasil
dan Pembahasan**

05

Kesimpulan

06

Daftar Pustaka



Pendahuluan



Covid-19 (Pandemi virus Korona) yang telah mengguncang seluruh dunia berasal dari Tiongkok, menyebabkan jumlah kematian melebihi 2,7 juta, lebih dari 127 juta orang yang terinfeksi penyakit ini, di seluruh dunia. (WHO, 2020). Sebagai akibat, praktik lockdown dan karantina yang telah diterapkan dunia membuat situasi panik dan stress pada masyarakat dunia (Sayema et al., 2021).

Dari permasalahan ini dilihat bagaimana perbandingan faktor-faktor sosial yang mempengaruhi stres di negara maju dan berkembang selama pandemi COVID-19 dapat diidentifikasi menggunakan metode klusterisasi dan regresi logistik dari data COVIDStress dunia.

Landasan Teori



Definisi Stress

Salah satu dampak dari pandemi Covid-19 adalah stres. Stres adalah suatu kondisi Ketika individu merasakan tekanan, baik secara tulus maupun secara intelektual. Tanda-tanda khas stres termasuk agitasi, kecemasan, dan lekas marah (Kemenkes, 2020).

Clustering

Clustering merupakan suatu metode untuk mengelompokkan dan mencari data dengan karakteristik yang memiliki kemiripan antara satu data dengan yang lain. Clustering memiliki dua jenis pengelompokan data, yaitu hierarchical dan non-hierarchical clustering. (Yuan, Chunhui, & Haitao Yang, 2019).

K-Prototypes

Algoritma K-Prototype adalah salah satu metode Clustering yang berbasis partitioning. Algoritma ini adalah hasil pengembangan dari algoritma K-Means (Huang,1998) untuk menangani clustering pada data dengan atribut bertipe campuran numerik dan kategorikal. Pada perhitungan jarak algoritma kprototype menggunakan ukuran kesamaan antar objek dengan menggabungkan persamaan eucdiean distance dengan dissimilarity measure yang ada dalam k-mode sebagai berikut:

$$d_{ij} = \sum_{k=1}^p (x_{ik} - x_{jk})^2 + \gamma \sum_{k=p+1}^{p+m} \delta(x_{ik}, x_{jk})$$

Secara umum algoritma K-Prototype terbagi kedalam tiga tahapan utama (Huang 1997), yaitu:

1. Inisialisasi awal prototype.
2. Alokasi objek di dalam X ke Cluster dengan prototype terdekat.
3. Realokasi object Jika terjadi perubahan prototype.

Klasifikasi

A decorative line graphic consisting of several connected line segments, starting from a black dot and extending towards the top right corner of the slide.

Klasifikasi merupakan suatu teknik yang digunakan untuk mengetahui atau memperkirakan kelas dari suatu objek berdasarkan atribut yang ada. Akan dilakukan analisis lebih lanjut terhadap negara-negara tersebut dengan memanfaatkan algoritma klasifikasi yaitu dengan menggunakan analisis regresi logistic.

2 rencana analisis yang akan dilakukan:

1. Mengetahui bagaimana kehidupan sosial seorang individu memengaruhi tingkat stres dari individu tersebut
2. Mengetahui faktor apa yang paling memengaruhi tingkat stres seorang individu secara umum

Regresi Logistik

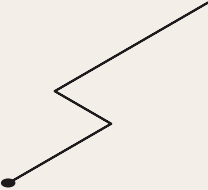
Regresi logistik adalah metode statistika yang digunakan untuk menganalisis hubungan antara satu atau lebih variabel independen dengan variabel dependen biner. Dalam regresi logistik, variabel dependen dianggap sebagai variabel acak yang mengikuti distribusi binomial

Langkah-langkah Regresi logistic:

1. **Pengumpulan Data:** data independen maupun dependen.
2. **Preprocessing Data**
3. **Pemilihan Model: Estimasi Model:** metode Maximum *Likelihood Estimation* (MLE).
4. **Uji Model:** uji goodness-of-fit, uji simultan, dan uji parsial.
5. **Interpretasi Model**

Metodologi

- **Preprocessing Data**
- **Clustering K-Prototypes**
- **Regresi Logistik**



Preprocessing Data



Mengganti Data Value dan Pemilihan Data yang dianalisis

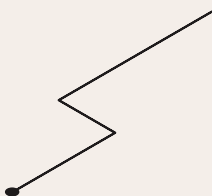
- Akan diganti data value pada 6 kolom yaitu: Scale_PSS10_UCLA_4, Scale_PSS10_UCLA_5, Scale_PSS10_UCLA_7, Scale_PSS10_UCLA_8, Compliance_4, Compliance_6.
- Menganalisis beberapa negara yang memiliki jumlah partisipan di atas nilai Q3 atau sebanyak 194 orang. Diperoleh ada sebanyak 42 negara yang memiliki jumlah partisipan di atas 194 orang. Negara-negara ini selanjutnya akan dikelompokkan menjadi negara maju dan negara berkembang untuk keperluan analisis lebih lanjut

Data Cleaning

- Missing Value
Membuang 20 kolom yang memiliki missing value dengan persentase di atas 70%, menggantikan nilai yang hilang dengan nilai yang ditentukan pada jenis data kategorik dan data numerik pada variabel demography.
- Feature Extraction
Menambahkan fitur baru yaitu dengan menghitung mean dari setiap variabel
- Outlier
mendeteksi outlier pada Expl_Distress, SPS, scale_SLON, Dem_age sebagai fitur yang akan dipilih dalam proyek ini

Reduksi Data

menghapus 8 kolom yang tidak akan diikutsertakan dalam proses analisis pada kolom yaitu : Duration..in.seconds., RecordedDate, UserLanguage, AD_gain, AD_loss, AD_check, Dem_Expatriate, Dem_state



Clustering Algoritma K-Prototypes

- Untuk clustering, digunakan beberapa variable yaitu umur, status isolasi, status pendidikan, status pekerjaan, status perkawinan, fitur 1 yang terdiri dari Expl_Distress_6, Expl_Distress_7, Expl_Distress_16, Expl_Distress_18, Expl_Distress_19, Expl_Distress_20 dan fitur 2 yang terdiri dari Expl_Distress_8, Expl_Distress_9, Expl_Distress_17, sps_mean, dan slon_mean.
- Didapatkan hasil cluster berikut dengan jumlah cluster = 3 :
- Hasil Cluster Negara Maju**

clusters	Dem_age_mean	SPS_mean	Slon_mean	feature_1_mean	feature_2_mean
1	49.51013514	4.881878378	2.189684685	2.670698198	1.937477477
2	42.58477039	5.65297423	1.693793193	2.049990312	1.450558677
3	32.66091682	5.352811909	2.963846881	3.530166982	2.305450536

clusters	Dem_employment_copy	Dem_gender_copy	Dem_maritalstatus_copy	Dem_edu	Dem_isolation_copy
1	Full time employed	Female	Married/cohabiting	College degree, bachelor, master	Life carries on with minor changes
2	Full time employed	Female	Married/cohabiting	College degree, bachelor, master	Life carries on with minor changes
3	Full time employed	Female	Married/cohabiting	College degree, bachelor, master	Life carries on with minor changes

- Hasil Cluster Negara Berkembang**

clusters	Dem_age_mean	SPS_mean	SLON_mean	feature_1_mean	feature_2_mean
1	51.93441754	4.998232279	2.287667197	3.03261446	2.419892758
2	30.04272695	5.150503018	3.193435121	3.911291277	3.199905314
3	35.68727467	5.554439845	1.947790092	2.268371389	1.850091245

clusters	Dem_employment	Dem_gender	Dem_maritalstatus	Dem_edu	Dem_isolation
1	Full time employed	Female	Married/cohabiting	College degree, bachelor, master	Life carries on with minor changes
2	Full time employed	Female	Single	College degree, bachelor, master	Isolated
3	Full time employed	Female	Married/cohabiting	College degree, bachelor, master	Life carries on with minor changes

Klasifikasi Regresi logistik



- Kelompokkan stress level menjadi 2 yaitu 0(low) dan 1(high).
0(low) batasnya adalah <2.5
1(high) batasnya adalah ≥ 2.5
- Standardisasi fitur yaitu:
dem_age,
sps_mean,
scale_slon_mean,
feature_1_maju_mean,
feature_2_maju_mean,
feature_1_berkembang_mean,
feature_2_berkembang_mean
- Lakukan regresi logistik
- Berdasarkan hasil cluster yang di dapat pisahkan negara berkembang menjadi 3 dataset begitu juga dengan negara maju
- Dataset yang didapat dibagi menjadi data train dan data test dengan rasio 70:30
- Data train dan data test dibagi lagi menjadi x_{train} , y_{train} , x_{test} , y_{test} .
 x_{train} x_{test} itu berisi data fitur , y_{train} y_{test} itu berisi data variable target

Hasil & Pembahasan



Negara Maju



- Hasil cluster 1 negara maju

```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(x_train_maju0, y_train_maju0)
eval_classification(logreg,x_train_maju0,x_test_maju0,y_train_maju0,y_test_maju0)

print(logreg.coef_)
print(logreg.intercept_)
```

Accuracy (Test Set): 0.7254469752632868
F1-Score (Test Set): 0.2032693674484719
Recall (Test Set): 0.12391681109185441
roc_auc (test-proba): 0.7016453968650316
roc_auc (train-proba): 0.6949354024319117
TP : 286
FP : 220
FN : 2022
TN : 5638

```
[[ 0.742789 -0.41296525  0.42243327 -0.34664131  0.35286038]]
[0.05932079]
```

- Hasil cluster 2 negara maju

```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(x_train_maju1, y_train_maju1)
eval_classification(logreg,x_train_maju1,x_test_maju1,y_train_maju1,y_test_maju1)

print(logreg.coef_)
print(logreg.intercept_)
```

➡ Accuracy (Test Set): 0.7537593984962406
F1-Score (Test Set): 0.850967007963595
Recall (Test Set): 0.9565217391304348
roc_auc (test-proba): 0.7347038871052585
roc_auc (train-proba): 0.7447329946852127
TP : 1870
FP : 570
FN : 85
TN : 135

```
[[ 0.62945921 -0.32861123  0.56675851 -0.44188959  0.37014643]]
[0.19744987]
```

- Hasil cluster 3 negara maju



```
from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(x_train_maju2, y_train_maju2)
eval_classification(logreg,x_train_maju2,x_test_maju2,y_train_maju2,y_test_maju2)

print(logreg.coef_)
print(logreg.intercept_)
```



```
Accuracy (Test Set): 0.7405991077119184
F1-Score (Test Set): 0.8421562924180725
Recall (Test Set): 0.9459812676976693
roc_auc (test-proba): 0.7315318255883172
roc_auc (train-proba): 0.7259521715243933
TP : 4343
FP : 1380
FN : 248
TN : 305

[[ 0.84094978 -0.4153587  0.40276208 -0.38630976  0.40206914]]
[0.03365965]
```

Negara Berkembang

- Hasil cluster 1 negara berkembang

```
▶ from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(x_train_berkembang0, y_train_berkembang0)
eval_classification(logreg,x_train_berkembang0,x_test_berkembang0,y_train_berkembang0,y_test_berkembang0)

print(logreg.coef_)
print(logreg.intercept_)
```

```
↳ Accuracy (Test Set): 0.68207343412527
F1-Score (Test Set): 0.7346791636625811
Recall (Test Set): 0.7880897138437741
roc_auc (test-proba): 0.7378788085173363
roc_auc (train-proba): 0.7531605415060647
TP : 1019
FP : 462
FN : 274
TN : 560

[[ 1.05111411 -0.27595076  0.23175433 -0.26654302  0.24248122]]
[0.86433888]
```

- Hasil cluster 2 negara berkembang

```
▶ from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(x_train_berkembang1, y_train_berkembang1)
eval_classification(logreg,x_train_berkembang1,x_test_berkembang1,y_train_berkembang1,y_test_berkembang1)

print(logreg.coef_)
print(logreg.intercept_)
```

```
↳ Accuracy (Test Set): 0.6501650165016502
F1-Score (Test Set): 0.5942936673625608
Recall (Test Set): 0.5678191489361702
roc_auc (test-proba): 0.7048704820678664
roc_auc (train-proba): 0.7089295806147484
TP : 854
FP : 516
FN : 650
TN : 1313

[[ 0.83067287 -0.31201317  0.41584967 -0.11768403  0.24294934]]
[0.69880911]
```


- Hasil cluster 3 negara berkembang

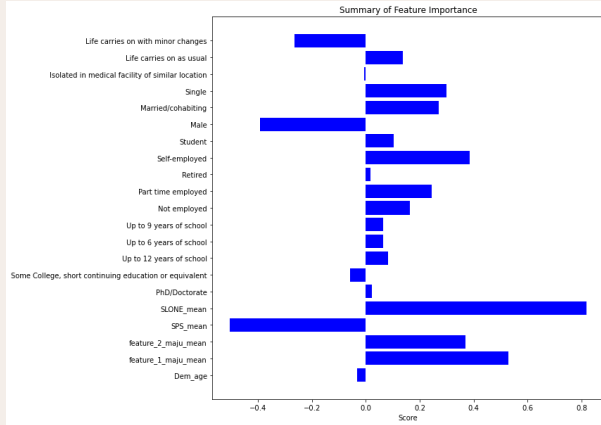
```
▶ from sklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(x_train_berkembang2, y_train_berkembang2)
eval_classification(logreg,x_train_berkembang2,x_test_berkembang2,y_train_berkembang2,y_test_berkembang2)

print(logreg.coef_)
print(logreg.intercept_)
```

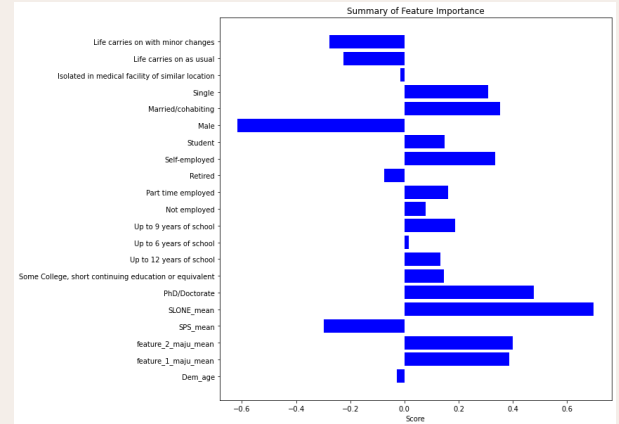
```
↳ Accuracy (Test Set): 0.8690880585314869
F1-Score (Test Set): 0.9299398685498531
Recall (Test Set): 0.9975997599759976
roc_auc (test-proba): 0.7232393279813811
roc_auc (train-proba): 0.7276235520882611
TP : 3325
FP : 493
FN : 8
TN : 1

[[ 1.02670557 -0.31833899  0.33357937 -0.17590166  0.27501277]]
[0.88844766]
```

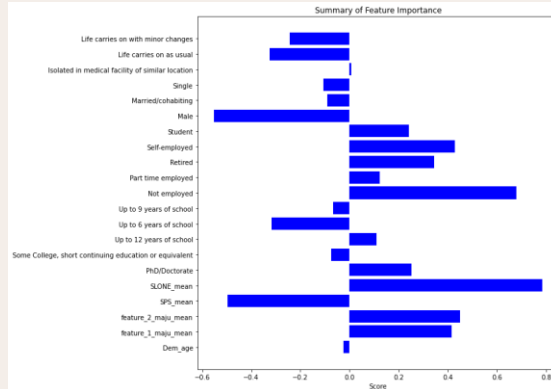
Feature Importance (Negara Maju)



Kluster 1

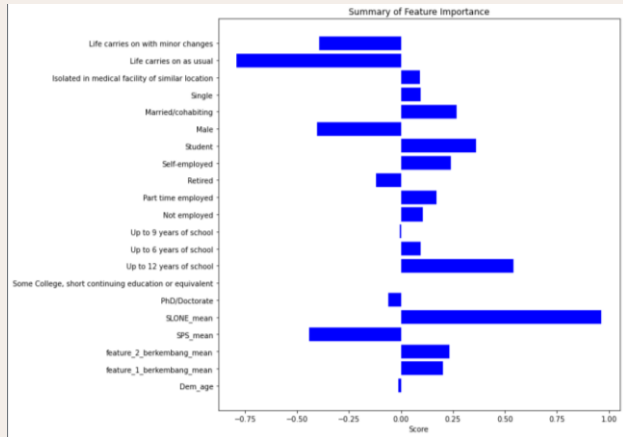


Kluster 2

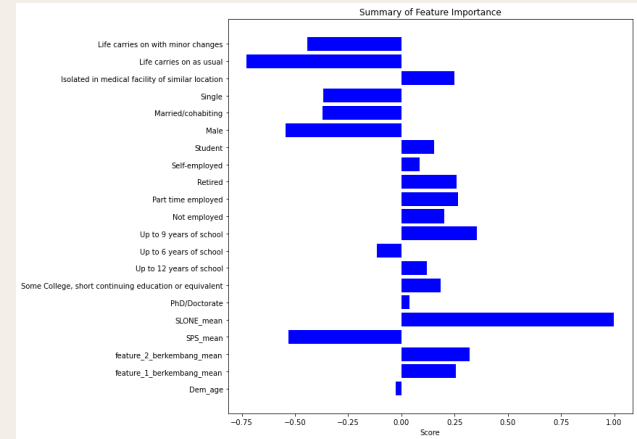


Kluster 3

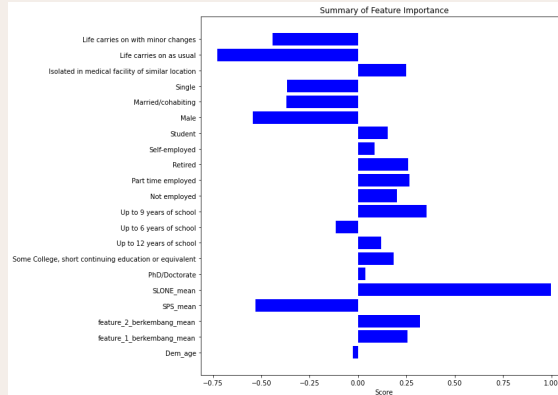
Feature Importance (Negara Berkembang)



Kluster 1



Kluster 2



Kluster 3

Kesimpulan




- a. Untuk hasil cluster negara maju skor stress tertinggi berada pada kelompok dengan rata rata umur 33, diikuti dengan kelompok dengan rata rata umur 50 dan 43. Setiap cluster pada negara maju didominasi oleh orang dengan karakteristik *full time employed*, *female*, *married/cohabiting*, *college degree*, *bachelor*, *master*, dan *life carries on with minor changes*.
- b. Untuk hasil cluster negara berkembang skor stress tertinggi berada pada kelompok dengan rata rata umur 30, diikuti dengan kelompok dengan rata rata umur 52 dan 36. Kluster 1 dan 3 pada negara berkembang didominasi oleh orang dengan karakteristik *full time employed*, *female*, *married/cohabiting*, *college degree*, *bachelor*, *master*, dan *life carries on with minor changes*. Sedangkan kluster 2, didominasi oleh orang dengan karakteristik *full time employed*, *female*, *single*, *college degree*, *bachelor*, *master*, dan *isolated*.
- c. Dari hasil Kluster juga dapat disimpulkan bahwa, untuk negara berkembang dan negara maju, ketika fitur 1, fitur 2, dan tingkat slon tinggi maka tingkat stress untuk kelompok tersebut juga akan tinggi.
- d. Untuk regresi logistik, dapat disimpulkan bahwa, model logistik biner, cukup baik untuk melakukan klasifikasi data untuk beberapa cluster.
- e. Variabel sosial seperti slon, fitur 1 yang terdiri dari Expl_Distress_6, Expl_Distress_7, Expl_Distress_16, Expl_Distress_18, Expl_Distress_19, Expl_Distress_20 dan fitur 2 yang terdiri dari Expl_Distress_8, Expl_Distress_9, Expl_Distress_17 merupakan fitur sosial yang memprediksi tingkat stress yang tinggi. Sedangkan fitur sosial SPS merupakan fitur sosial yang memprediksi Tingkat stress yang rendah.

Daftar Pustaka

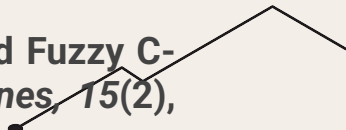


1. Yuan, Chunhui, and Haitao Yang. "Research on K-value selection method of K-means clustering algorithm." J 2.2 (2019): 226-235.
2. Agresti, A. 2002. Categorical Data Analysis, John Wiley and Sons, Inc. New York.
3. Ayungtyas, D. A. (2017). Klasifikasi Menggunakan Metode Regresi Logistik dan Support Vector Machine. Skripsi. Universitas Brawijaya, Malang.
4. Chairudin, H. B., Asrifuddin, A., & Ratag, B. T. (2023). Faktor-Faktor yang Berhubungan dengan Tingkat Stres Masyarakat Pasca Pandemi COVID-19 di Desa Tompaso II Kabupaten Minahasa. *Jurnal Kesehatan Tambusai*, 4(2), ISSN: 2774-5848 .
<https://journal.universitaspahlawan.ac.id/index.php/jkt/article/view/15896/12513>
5. Everitt, B. S., Landau, S., Leese, M., & Stah, D. (2011). Cluster Analysis. Wiley Series in Probability and Statistics. King's College London, UK. ISBN 978-0-470-74991-3.
https://cicerocq.files.wordpress.com/2019/05/cluster-analysis_5ed_everitt.pdf
6. Feizi, A., Aliyari, R., & Roohafza, H. (2012). Association of Perceived Stress with Stressful Life Events, Lifestyle and Sociodemographic Factors: A Large-Scale Community-Based Study Using Logistic Quantile Regression. *Computational and Mathematical Methods in Medicine*, 2012, 151865. <https://doi.org/10.1155/2012/151865>

7. Ganmanah, M., & Kudus, A. (2020). Penerapan Algoritme K-Prototypes untuk Pengelompokan Desa-Desa di Provinsi Jawa Barat Berdasarkan Indikator Indeks Desa Membangun Tahun 2020. *Prosiding Statistika*, Volume 0(Issue 0), Halaman 543. DOI: <http://dx.doi.org/10.29313/v0i0.28974>
8. Gupta, A., Gupta, A., & Mishra, A. (2016). Research Paper on Cluster Techniques of Data Variations. *International Journal of Advance Technology & Engineering Research (IJATER)*, 1(1), 39. ISSN 2250-3536. : <https://www.researchgate.net/publication/265077297>
9. Hosmer D. W and S. Lemeshow, 2000. *Applied Logistic Regression*. John Wiley and Sons, New York
10. Huang, Z. (1997). Clustering Large Data Sets with Mixed Numeric and Categorical Values. *Computer Science, Mathematics*. Corpus ID: 3007488
11. Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large data Sets with Categorical Values . *Data Mining and Knowledge Discovery*, 2833-304.
12. Islam, S.D.-U., Bodrud-Doza, M., Khan, R.M., Haque, M.A., Mamun, M.A., 2020. Exploring COVID-19 stress and its factors in Bangladesh: a perception-based study. *Heliyon* 6 (7), e04399
13. Kementerian Kesehatan Republik Indonesia (Kemenkes RI). (2020). *Situasi Terkini Perkembangan Novel Coronavirus (COVID-19)*

- 
14. Le, H.T., Lai, A.J.X., Sun, J., Hoang, M.T., Vu, L.G., Pham, H.Q., Le, X.T.T., 2020a. Anxiety and depression among people under the nationwide partial lockdown of Vietnam. *Front. Public Health* 8, 656
 15. Luo, Y., Chua, C.R., Xiong, Z., Ho, R.C., Ho, C.S., 2020. A systematic review of the impact of viral respiratory epidemics on mental health: an implication on the coronavirus disease 2019 pandemic. *Front. Psychiatr.* 11
 14. Mamun, M.A., Akter, T., Zohra, F., Sakib, N., Bhuiyan, A.I., Banik, P.C., Muhit, M., 2020. Prevalence and risk factors of COVID-19 suicidal behavior in Bangladeshi population: are healthcare professionals at greater risk? *Heliyon* 6 (10), e05259
 14. Nafisah, Q., & Chandra, N. E. (2017). Analisis Cluster Average Linkage Berdasarkan Faktor-Faktor Kemiskinan di Provinsi Jawa Timur. *Zeta – Math Journal*, 3(2). ISSN: 2459-9948.
 15. Ripon, R.K., Mim, S.S., Puente, A.E., Hossain, S., Babor, M.M.H., Sohan, S.A., Islam, N., 2020. COVID-19: psychological effects on a COVID-19 quarantined population in Bangladesh. *Heliyon* 6 (11), e05481

16. Syafiyah, U., Asrafi, I., Wicaksono, B., Puspitasari, D. P., & Sirait, F. M. (2022). Analisis Perbandingan Hierarchical dan Non-Hierarchical Clustering Pada Data Indikator Ketenagakerjaan di Jawa Barat Tahun 2020. Seminar Nasional Official Statistics. Halaman 803. <https://prosiding.stis.ac.id>
17. Serafini, G., Parmigiani, B., Amerio, A., Aguglia, A., Sher, L., Amore, M., 2020. The psychological impact of COVID-19 on the mental health in the general population. QJM: Int. J. Med. 113 (8), 531–537.
18. Sher, L., 2020. The impact of the COVID-19 pandemic on suicide rates. QJM: Int. J. Med. 113 (10), 707–712.
19. Sultana, S., Shafique, I., Majeed, N., Jamshed, S., Shahani, A. K., & Qureshi, F. (2021). Impact of Covid-19 outbreak on psychological health–The case of Bangladesh. *Heliyon*, 7, e06772. Diakses dari <https://www.cell.com/heliyon>

- 
20. Wulandari, L., & Yogantara, B. O. (2022). Algorithm Analysis of K-Means and Fuzzy C-Means for Clustering Countries Based on Economy and Health. *Journal unnes*, 15(2), 109-116. DOI: <https://doi.org/10.30998/faktorexacta.v15i2.12106>
 21. World Health Organization, 2020a. Coronavirus (COVID-19). Retrieved from. <https://covid19.who.int/>.
 23. Zubayer, A.A., Rahman, M.E., Islam, M.B., Babu, S.Z.D., Rahman, Q.M., Bhuiyan, M.R.A.M., Habib, R.B., 2020. Psychological states of Bangladeshi people four months after the COVID-19 pandemic: an online survey. *Heliyon* 6 (9), e05057