Data Mining Final Project

"The Effect of Covid-19 Cases to Change in Transit and Residential Category in Google Mobility Data in Turkey"

Feyza Becer- 19120205036, Yusuf Özkan- 19120205038

1) Implementation and Source Code

In this project, we used the Python programming language for data preprocessing, analysis and prediction. The main Python packages used were Pandas, Numpy, Matplotlib, Scikit-Learn, Seaborn and Plotly. Pandas and Matplotlib were used for data pre-processing as well as visualization. Scikit-Learn was used for K-Means Clustering and Random Forest Classification. Seaborn and Plotly were used for data visualization.

Our implementation can be found in our GitHub repository at https://github.com/ysfzkn/covid_mobility_data_analysis

2) Data Set and Data Processing Details

We used the change Transit Stations and Residential columns in Google Mobility Data for Turkey, years 2020 and 2021. And we created a dataset with provided Covid-19 daily new cases data by the Ministry of Health of Turkey. We analyzed the data on a daily and weekly basis. We detected many outliers daily analysis cause of the lockdown on weekends in Turkey. Therefore, we analyze the data on weekly basis for the year 2020 for clustering and prediction.

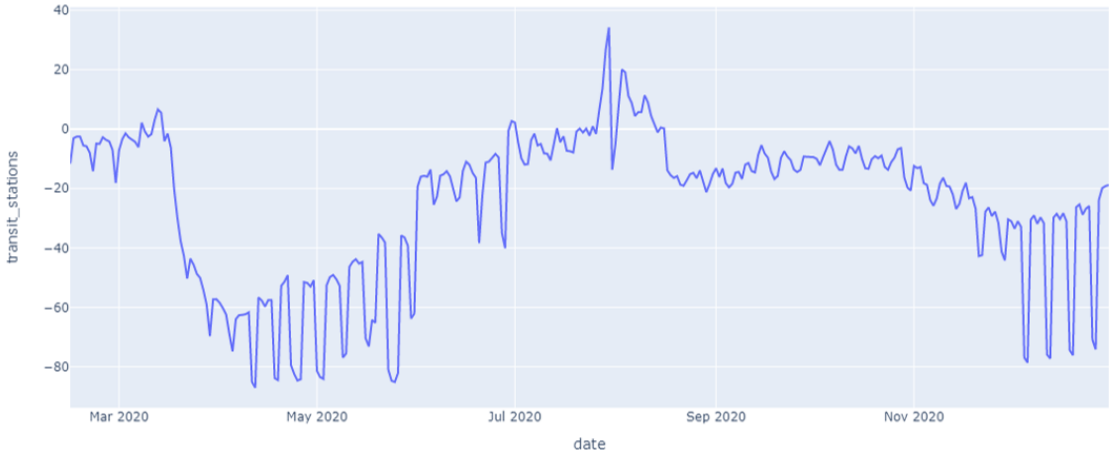

Figure 1: Line for 2020 Change in Transit Stations.

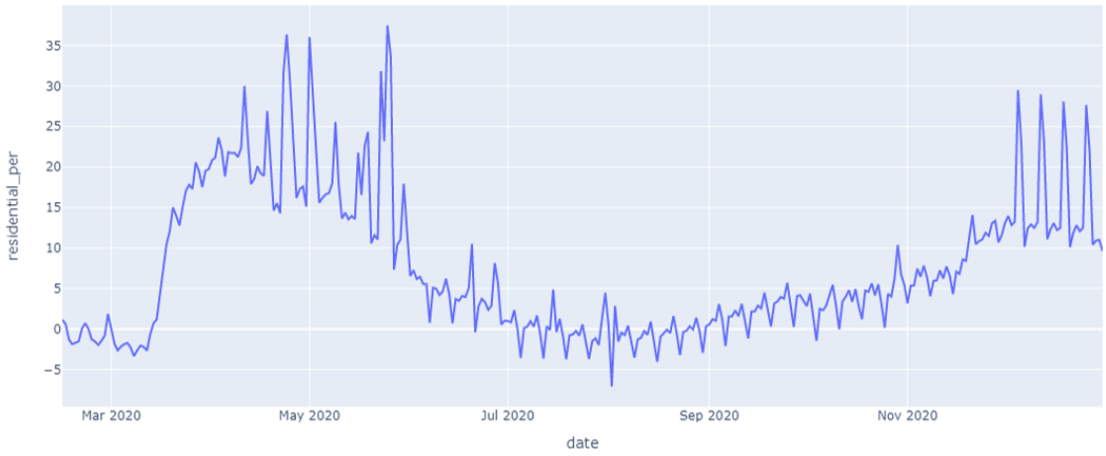

Figure 2: Line for 2020 Change in Residential.

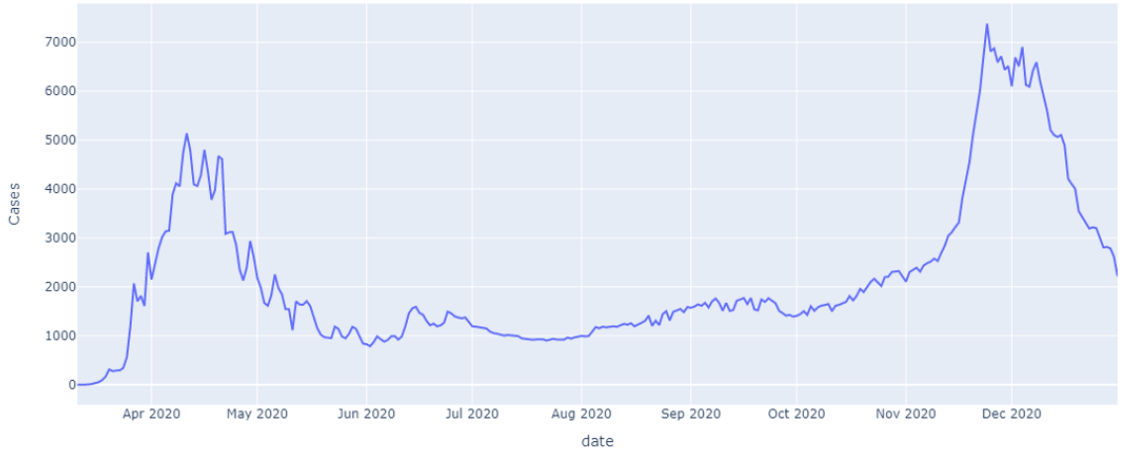

Figure 3: Line for 2020 New Cases.

3) K-Means Clustering

Clustering is a machine learning method that is typically used to group a set of objects into different clusters. We used K-Means Clustering in this project. The main purpose is to understand the relationship between cases, transit stations, and residential data separately. When choosing to k value for cluster count, we used the Elbow Method. According to Elbow Method, we set the k value as 3. We got a more explainable graph for the weekly basis data than for the daily basis data. So, we have been selected the weekly basis data to use. We clustered the transit data and residential data separately and visualized them with two separate scatter plots.

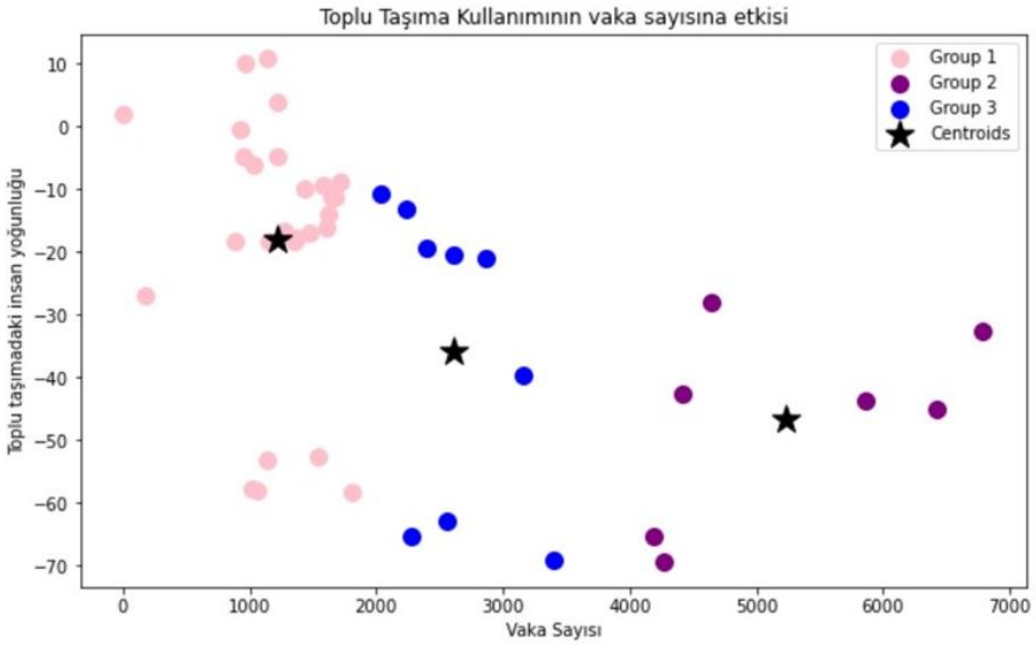

Figure 4:

On the x-axis, we have the new cases data and the y-axis represents the change in transit stations. The centroids represent by stars.

We have 3 clusters that Group 1, 2 and 3 as shown.

Group 1; When the number of cases was around 1000, people used public transportation stations and there was a decrease of around 10%, but there was no clear decrease.

Group 2; When the number of cases was the highest, we can see that people have reduced their use of public transportation and this decrease has been gathered around 50%. Also, this group has the most decrease rate at around 70%.

Group 3; We can see that the centroid is lower than group 1. While the number of cases was at an average level, we can see that people tend to reduce their use of public transport even more.

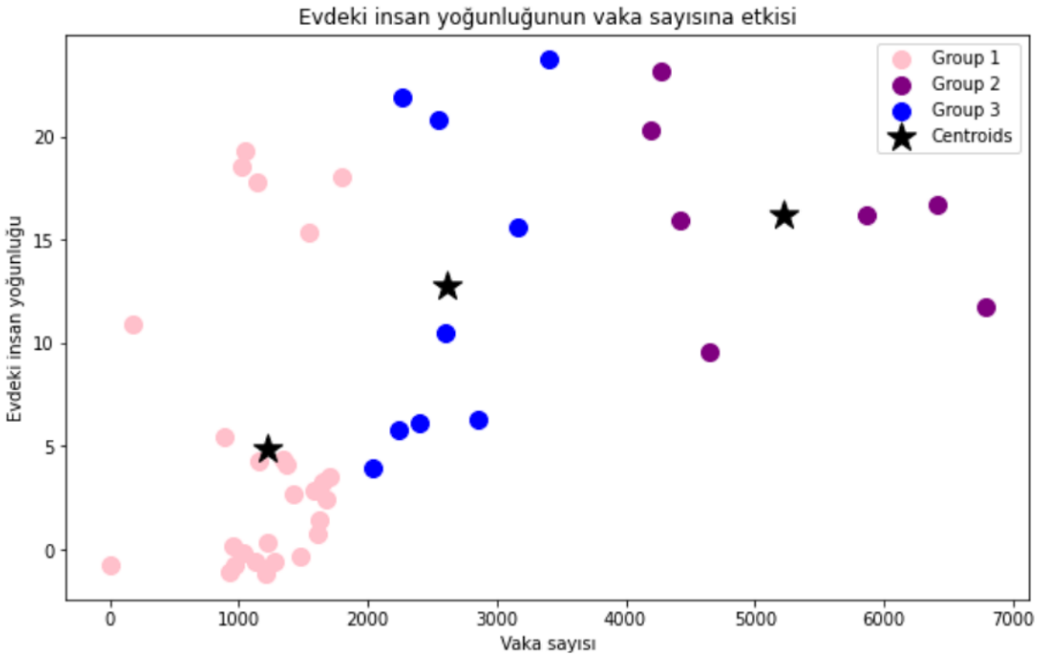

Figure 5:

On the x-axis, we have the new cases data and the y-axis represents the change in residential. The centroids represent by stars.

We have 3 clusters that Group 1, 2 and 3 as shown.

Group 1; We observed that the data are concentrated in places where the rate of staying at residential has not increased or has increased slightly, except for a few outliers.

Group 2; When the number of cases is highest, although we have scattered data, we can clearly see that people were preferred to stay at residential around approximately rate 15%.

Group 3; When the number of cases is average, we can see people were preferred to stay at residential around rate 12%. Also, we have high rates in this group. We have researched and examined the data for these outliers; We have seen that according to the news by the Ministry of Health in Turkey when the number of cases was about is average there is a lockdown on these dates. Consequently, we think these outliers are because of the lockdowns.

4) Forecasting with Random Forest

Random Forest is a machine learning technique that's used to solve regression and classification problems. We used Random Forest Regression in this project.

Using this algorithm, we have trained the model with the number of cases and predicted the transit and residential values. Before implementing a random forest, we separated the data as training and test by date.

We calculated the R^2 score from the difference between the predicted and actual values. R^2 score provides a measure of how well future samples are likely to be predicted by the model.

For transit stations, we calculated the R^2 score as `80.14%`.

For residential we calculated the R^2 score as `83.09%`.

We visualized the real transit stations and residential values with predicted values on a date basis.

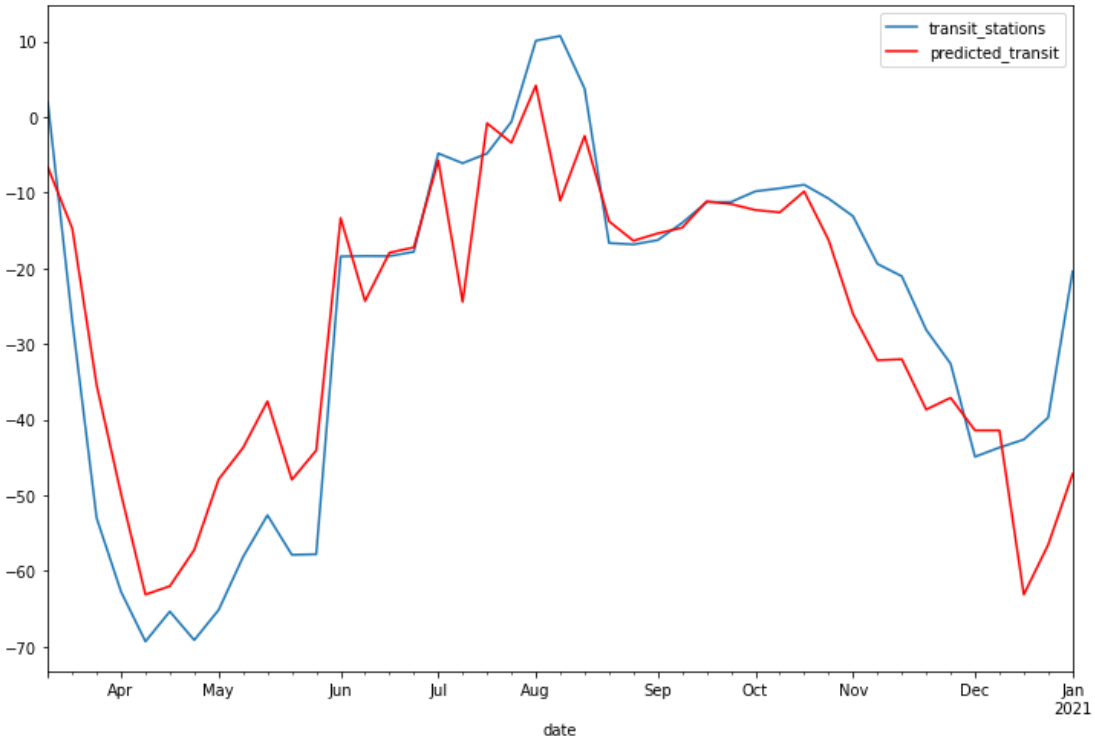

Figure 6: Plot for the change in transit stations and predicted values on a date basis.

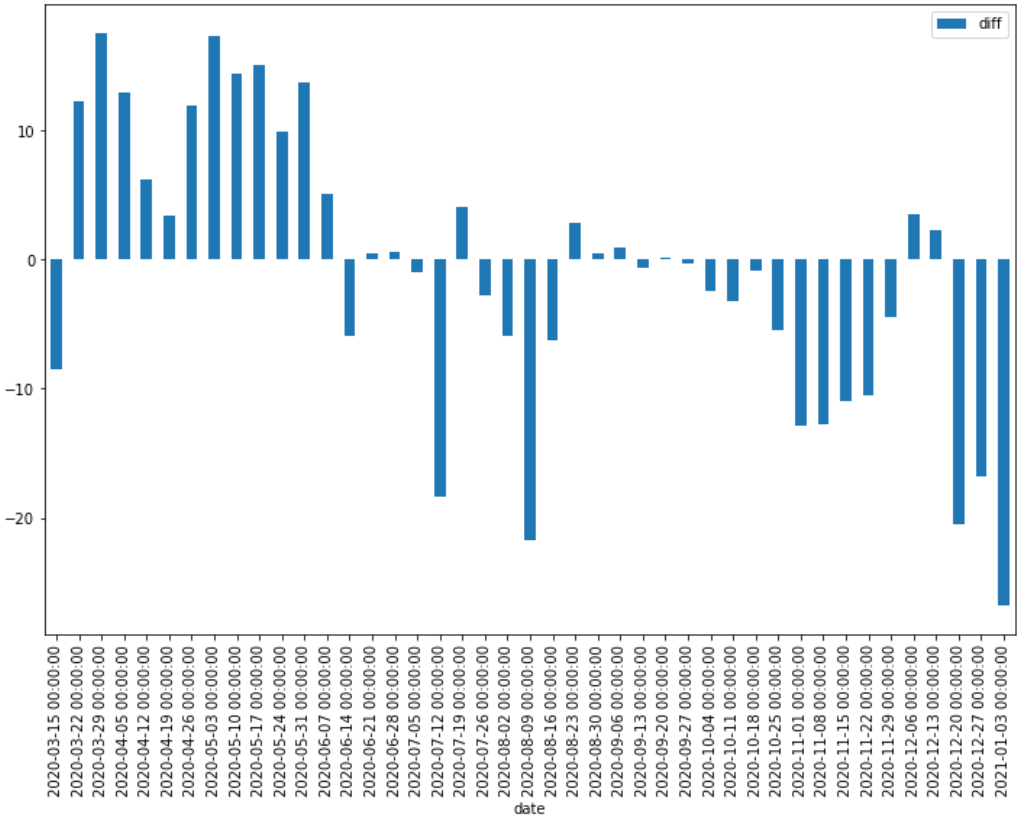

Figure 7: Bar plot for the difference between actual values of train stations and predicted values.

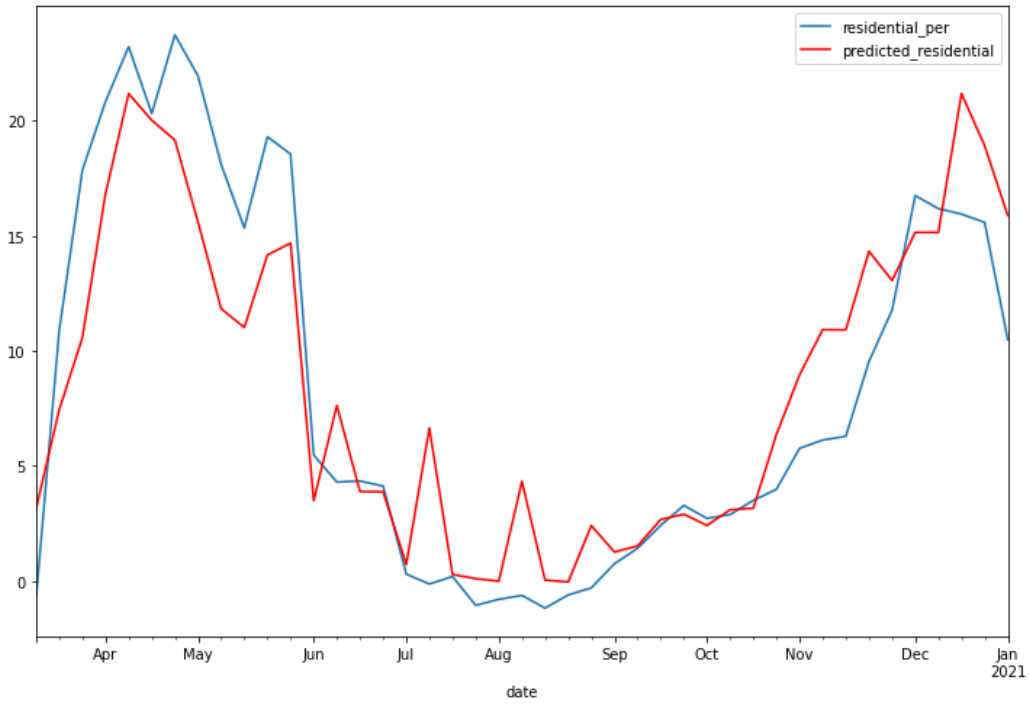

Figure 8: Plot for the change in residential and predicted values on a date basis.

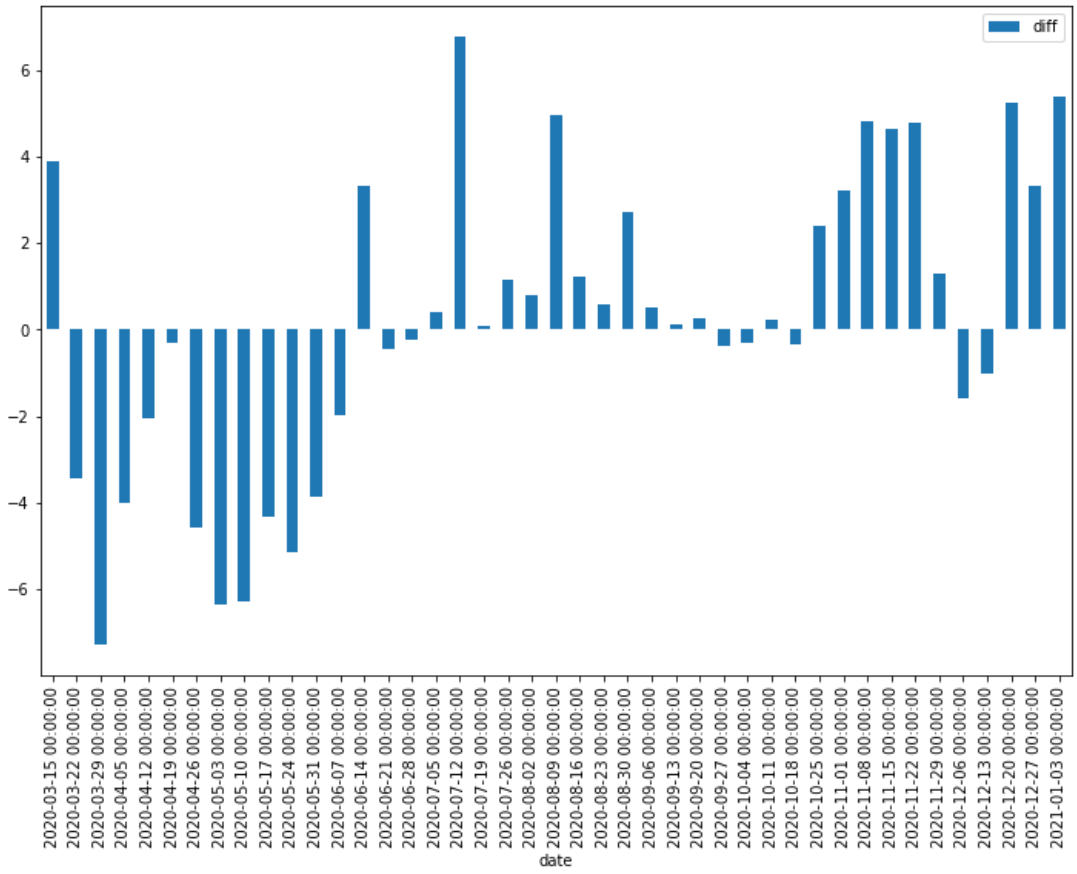

Figure 9: Bar plot for the difference between actual values of residential and predicted values.

We realized that we have not enough training data for the more accurate scores. Although, we can see that predicted values similar to actual ones for both of data.

Consequently, we observed the clusters of change in transit stations and residential data from Google Mobility by New Cases data in Turkey. As a result, we preprocessed the data and applied K-Means Cluster Analysis by the number of cases. Looking at the cluster analysis, we can say that people generally took the number of cases into account. When the number of cases is high, we can say that the mobility in the transit stations decreases and the mobility in the residential increases.

In addition, we made predictions for future dates with the Random Forest Regression algorithm using these case numbers.

References

[1] *Raval, U.R. and Jani, C. (2016) Implementing & Improvisation of K-Means Clustering Algorithm. International Journal of Computer Science and Mobile Computing, 5, 191-203.*

[2] *Marutho, D., Handaka, S.H. and Wijaya, E. (2018, September) The Determination of Cluster Number at K-Mean Using Elbow Method and Purity Evaluation on Headline News. 2018 International Seminar on Application for Technology of Information and Communication, Semarang, 21-22 September 2018, 533-538.*

[3] *Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston*

[4] *MacKay DJC. Information Theory, Inference & Learning Algorithms. Cambridge University Press; 2002.*

[5] *https://scikit-learn.org/stable/supervised_learning.html#supervised-learning*