PREDICTION OF GLUCOSE LEVEL IN DIABETICS WITH SUPPORT VECTOR REGRESSION

Devi Wulandari[1*]; Agus Subekti[1#,2]

Masters in Computer Science
[1, 2] Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri
[1]www.nusamandiri.ac.id
[1]dw91325@gmail.com; [1#]agus@nusamandiri.ac.id

[2]Pusat Penelitian Informatika LIPI
[2]Lembaga Ilmu Pengetahuan Indonesia (LIPI)
[2]http://informatika.lipi.go.id/
[2]agus.subekti@lipi.go.id

* Corresponding Author

Abstract— One of the common diabetes factors that people hear is that they consume too much or often consume sweet foods or drinks so that blood sugar in the human body increases. The times and increasingly sophisticated technology make it easier for someone to be able to predict a disease such as diabetes with machine learning techniques. Therefore, from the existing problems, a machine learning technique will be made in predicting glucose levels in diabetics. The aim is to predict glucose levels in diabetics and find the best algorithm from several comparison algorithms. The results of the experiments carried out by the support vector regression algorithm have a lower mean squared error value of 28.9480 compared to other comparative algorithms and visualize the error classification seen that Instance no 47 has a prediction of the highest plasma glucose value of 189.2305.

Keywords: Diabetes, Glucose Level, Support Vector Regression

Abstrak*— Salah satu faktor penyakit diabetes yang biasa di dengar oleh masyarakat yaitu karena terlalu banyak atau seringnya mengkonsumsi makanan atau minuman manis sehingga gula darah pada tubuh manusia meningkat. Perkembangan zaman serta teknologi yang semakin canggih mempermudah seseorang untuk dapat memprediksi suatu penyakit salah satunya diabetes dengan teknik machine learning. Oleh karena itu, dari permasalahan yang ada akan dibuatkan teknik machine learning prediksi level glukosa pada penderita diabetes. Tujuannya adalah untuk memprediksi level glukosa pada penderita diabetes dan menemukan algoritma terbaik dari beberapa algoritma pembanding. Hasil dari eksperimen yang dilakukan algoritma support vector regression memiliki nilai root mean squared arror lebih rendah yaitu 28,9480 dibanding dengan algoritma yang pembanding yang lain dan visualize clasifier error dilihat bahwa Instance no 47memiliki prediksi nilai plasma glukosa tertinggi yaitu 189,2305.*

Kata Kunci*: Diabetes, Level Glukosa, Support Vector Regression*

INTRODUCTION

Diabetes is a disease that we have been very common to hear and meet. Some factors that cause diabetes are emerging such as genetic factors, weight loss, food and bad habits in daily life (Soumya & Srilatha, 2011). Machine learning with the development of algorithms and techniques (Kaur & Kumari, 2018) with Artificial Intelligence (AI) intelligence can identify and understand each data input so that it can predict results and make decisions (Sun, 2013).

One of the common diabetic factors that people hear is that they consume too much or often eat sweet foods or drinks so that blood sugar in the human body increases (Anggraini, 2019). At this time blood sugar checks to determine the level of sugar in the human body have often been done in health centers or hospitals. However, the development of the times and technology have a lot of researchers doing research on diabetes with machine learning techniques.

Neural Network models provide a suitable construction for glucose prediction in which many factors influence and are indicators of future glycemic trends (Pappada et al., 2011)(Alloghani et al., 2019). Given the increasing use of CGM technology, NNMs trained using CGM data have

DOI: 10.33480/pilar.v16i1.1264

become the focus of investigations recently (Pappada et al., 2011). Meanwhile according to diabetes mellitus is a global problem (Robertson et al., 2011) that continues to increase. However, it has been shown that the set is good Blood glucose levels (BGL) (Robertson et al., 2011). Related complications and expensive costs can be reduced significantly. In this learning example, Elman repeated Artificial Neural Networks (ANNs) used to make BGL predictions (Sandham et al., 2011) based on BGL history (Al-Khasawneh & Hijazi, 2014), food intake, and insulin.

This research only discusses the prediction of diabetes in female patients with 8 (eight) independent variables or attributes, namely pregnancy, plasma glucose, diastolic blood pressure, triceps skin fold, serum insulin, body-mask index, diabetes pedigree and age (Sarojini et al., 2009) and 1 dependent variable namely the diabetes and non-diabetes class label. Then, plasma glucose variable is used as a factor that influences whether or not patients are tested using a regression method, namely support vector regression, Neural Network, Random Forest, K-Nearest Neighbors (KNN).

The purpose of this study is to predict glucose levels in diabetics with the research methodology used in this research is to use the Knowledge Discovery in Database (KDD) (Kavakiotis et al., 2017).

MATERIALS AND METHODS

Lack of insulin (insulin resistance) will cause diabetes. The main energy source in the human body is glucose which is used for all metabolic processes. The regulation of glucose movement in the body is regulated by hormones from the pancreas. As a result of diabetes affects the body's metabolic disorders, and complications arising from impaired kidney function, impaired eye function, disorders of blood vessels, and other related complications.. The complications that occur as a result of diabetes will affect the body's metabolism. This diabetic disease often impacts on blindness, causes heart disease, stroke, kidney failure, and so forth. Some things from microvascular which are attenuated into complications by diabetes, including nephropathy, retinopathy, and neuropathy, as well as atherosclerosis and stroke (Soumya & Srilatha, 2011).

Blood sugar or glucose is the amount of glucose that can be found in the bloodstream of humans which generally have units of mg / dL. The body naturally regulates blood sugar levels with the help of the hormone insulin produced by pancreatic beta cells as part of the body's homeostasis(Belinda, 2019). Blood sugar levels can change at any time, especially before and after eating and when the body will rest or sleep. This happens because after eating, the digestive system will begin to break down carbohydrates into sugar or glucose so that it can be absorbed by the body and processed into energy(Belinda, 2019)

Support Vector Machine (SVM) is a vector-based vector model that requires a text to be converted into a vector before it is used for classification. The key idea of SVM is to find (Waila et al., 2012) the maximum decision surface (hyperlane) of each data point to conduct training machines supported by vectors or commonly called SVM that require a very large Quadratic Programming (QP) solution (Saputra et al., 2016). Regression is a process that can make a prediction of various patterns that have been previously formed as a data model. The purpose of regression is to create a new variable that represents a representation of data development in the future. WEKA supports the regression process and this is facilitated with a simple user interface or user experience.

In this study, the type used is Absolute Experiment where in this research predict glucose levels in diabetics with support vector regression with the population in this study are diabetics and the sample used in this study is the data of female patients with age ranging from 21 years old who has diabetes.

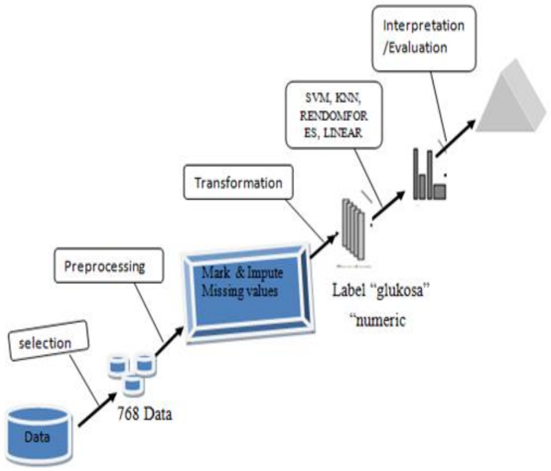
The dataset for this study is the Indian Pima population taken from repositories https://archive.ics.uci.edu/ml/datasets/diabetes which contains data on female patients with a minimum age of 21 years with a total of 8 attributes and 768 instances and classified into 2 (two) classes, namely diabetes and non-diabetes.

Table 1 Diabetes of Indian Pima dataset

No. Attribute	Attribute	Variable Type	Distance
A1	Pregnancy	Integer	0-17
A2	Plasma Glucose	Real	0-199
A3	Diastolic Blood Pressure	Real	0-122
A4	Triceps Skin Fold	Real	0-99
A5	Serum Insulin	Real	0-846
A6	Body Mass Index	Real	0-67,1
A7	Diabetes Pedigree	Real	0,078-2,42
A8	Age	Integer	21-81
Class		Binary	1= positive diabetes 0 = negative diabetes

Sources:(Kaur & Kumari, 2018)

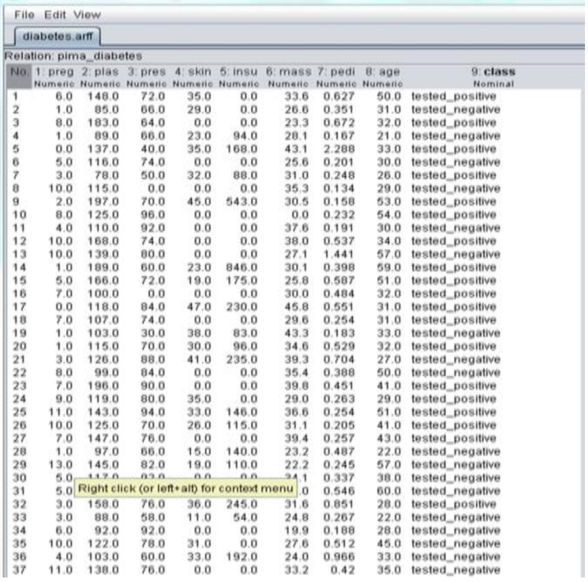
In this study the data collection method used is secondary data where the main data obtained from the UCI machine learning repository with a total of 768 data classified into 2 (two) classes, namely diabetes and non-diabetes with 8 (eight) risk factors namely number of pregnancies, glucose plasma, blood pressure, skinfold thickness, insulin, weight, diabetes pedigree and age.



Sources: (Wulandari & Subekti, 2019)
Figure 1 Steps in the Knowledge Discovery Process in Database (KDD)

RESULTS AND DISCUSSION

Data taken through the Pima Indian dataset contained in the UCI repository with 768 instances (8 attributes and 1 label), here are some original diabetes datasets:



Relation: pima_diabetes

No.	1: preg Numeric	2: plas Numeric	3: pres Numeric	4: skin Numeric	5: insu Numeric	6: mass Numeric	7: pedi Numeric	8: age Numeric	9: class Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	tested_positive
2	1.0	85.0	66.0	29.0	0.0	26.6	0.351	31.0	tested_negative
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	tested_positive
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	tested_negative
5	0.0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	tested_positive
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	tested_negative
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	tested_positive
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	tested_negative
9	2.0	197.0	70.0	45.0	543.0	30.5	0.158	53.0	tested_positive
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	tested_positive
11	4.0	110.0	92.0	0.0	0.0	37.6	0.191	30.0	tested_negative
12	10.0	168.0	74.0	0.0	0.0	38.0	0.537	34.0	tested_positive
13	10.0	139.0	80.0	0.0	0.0	27.1	1.441	57.0	tested_negative
14	1.0	189.0	60.0	23.0	846.0	30.1	0.398	59.0	tested_positive
15	5.0	166.0	72.0	19.0	175.0	25.8	0.587	51.0	tested_positive
16	7.0	100.0	0.0	0.0	0.0	30.0	0.484	32.0	tested_positive
17	0.0	118.0	84.0	47.0	230.0	45.8	0.551	31.0	tested_positive
18	7.0	107.0	74.0	0.0	0.0	29.6	0.254	31.0	tested_positive
19	1.0	103.0	30.0	38.0	83.0	43.3	0.183	33.0	tested_negative
20	1.0	115.0	70.0	30.0	96.0	34.6	0.529	32.0	tested_positive
21	3.0	126.0	88.0	41.0	235.0	39.3	0.704	27.0	tested_negative
22	8.0	99.0	84.0	0.0	0.0	35.4	0.388	50.0	tested_negative
23	7.0	196.0	90.0	0.0	0.0	39.8	0.451	41.0	tested_positive
24	9.0	119.0	80.0	35.0	0.0	29.0	0.263	29.0	tested_positive
25	11.0	143.0	94.0	33.0	146.0	36.6	0.254	51.0	tested_positive
26	10.0	125.0	70.0	26.0	115.0	31.1	0.205	41.0	tested_positive
27	7.0	147.0	76.0	0.0	0.0	39.4	0.257	43.0	tested_positive
28	1.0	97.0	66.0	15.0	140.0	23.2	0.487	22.0	tested_negative
29	13.0	145.0	82.0	19.0	110.0	22.2	0.245	57.0	tested_negative
30	5.0	117.0	92.0	0.0	0.0	34.1	0.337	38.0	tested_negative
31	5.0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.546	60.0	tested_negative
32	3.0	158.0	76.0	36.0	245.0	31.6	0.851	28.0	tested_positive
33	3.0	88.0	58.0	11.0	54.0	24.8	0.267	22.0	tested_negative
34	6.0	92.0	92.0	0.0	0.0	19.9	0.188	28.0	tested_negative
35	10.0	122.0	78.0	31.0	0.0	27.6	0.512	45.0	tested_negative
36	4.0	103.0	60.0	33.0	192.0	24.0	0.966	33.0	tested_negative
37	11.0	138.0	76.0	0.0	0.0	33.2	0.42	35.0	tested_negative

Sources: (Kahn, 1994)

Figure 2 Indian Pima Diabetes dataset

Can be seen that in some of the diabetes data that has not been processed in the picture above there are some data with a value of 0 which means that the data is missing values so that this data will go through preprocessing so that there is no data with a value of 0 or missing values .

Based on the diabetes dataset above, there are some data with 0 values or missing values, so it needs to be processed so that there are no more data with missing values.

One problem that often occurs in research is missing data or commonly known as missing data (Santoso. S, 2012). Missing data is information that is not available for an object. There are several ways that can be done to deal with missing data(Mukarromah et al., 2015) such as: Listwise deletion, Pairwise deletion and Imputation. Listwise deletion is to delete cases (objects) that contain missing data. Pairwise deletion is by removing missing data, so that only the available values are analyzed. Imputation is to fill in missing data with possible values based on information available on the data.

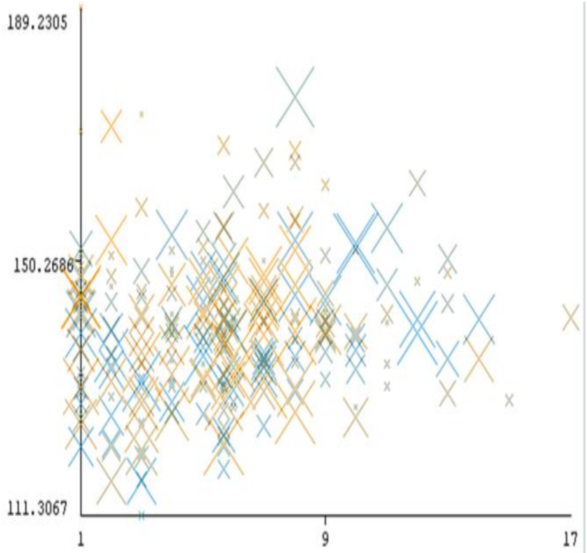
After completing the preprocessing stage and changing the glucose level into a major and influential factor, the results obtained from this study by conducting several algorithm experiments, namely:

Table 2 Summary of Diabetes Positive Data expected in a Support Vector Regression

SVM	N	TN
Correlation coefficient	0.241	0.2413
Mean absolute error	23,8534	23,8554
Root mean squared error	28,9548	28,9480
Relative absolute error	95,5132%	95,5210 %
Root relative squared error	98,2801%	98,2569 %
Total Number of Instances	268	268

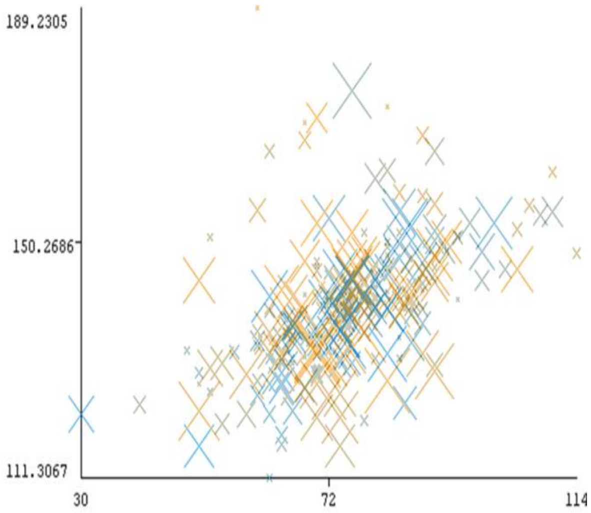
Sources: (Wulandari & Subekti, 2019)

Seen from table 2 diabetes data results using the SVM algorithm with 2 (two) experiments, namely by normalizing the data and without normalizing. It is seen that the lowest Root Mean Squared Error (RMSE) value is found in experiments with un normalized data with a value of 28.9460.



Sources: (Wulandari & Subekti, 2019)
Figure 3: Plot prediction of plasma glucose in pregnancy

Figure 3 shows the plot or visualize error class between X variable, namely preggnancy (pregnancy) and Y variable, which is prediction of plasma glucose with values ranging from 1 to 17 and x values ranging from 111.3067 to 189.2305, indicating that there are more errors in pregnancy between 1-9 with a plasma predictive value of 189,2305 mg.dl.



Sources: (Wulandari & Subekti, 2019)
Figure 4 Plot prediction of plasma glucose in blood pressure

Figure 4 shows the plot or visualize error classifier between variable X, namely blood pressure (blood pressure) with a minimum value of 30 mm hg and a maximum of 114 mm hg and Y variable, prediction of plasma glucose 111.3067 to 189.2305.

Table 3 Results of Positive Diabetes Data with Neural Network

Neural Network	N	TN
Correlation coefficient	0.1404	0.1404
Mean absolute error	28,7590	28,7590
Root mean squared error	35,5662	35,5662
Relative absolute error	115,1565 %	115,1565 %
Root relative squared error	120,7207 %	120,7207%
Total Number of Instances	268	268

Sources: (Wulandari & Subekti, 2019)

Based on table 3 the results of non-diabetic data using Neural Network algorithm with 2 (two) experiments, namely with data normalization and without normalization. It was seen that the lowest Root Mean Squared Error (RMSE) value was found in experiments with normalized data and without normalization with a value of 35.5662.

Table 4 Results of Positive Diabetes Data with Random Forest

Random Forest	N	TN
Correlation coefficient	0.2818	0.2953
Mean absolute error	23,4076	23,3917
Root mean squared error	28,5091	28,3303
Relative absolute error	93,7285 %	93,6645 %
Root relative squared error	96,7673 %	96,1604%
Total Number of Instances	268	268

Sources: (Wulandari & Subekti, 2019)

Table 4 shows the results of positive diabetes data using the random forest algorithm with 2 (two) experiments, namely data normalization and without normalization. It is seen that the lowest Root Mean Squared Error (RMSE) value is in experiments with data without normalization with a value of 28.3303.

DOI: 10.33480/pilar.v16i1.1264

Table 5 Results of Positive Diabetes Data with KNN

KNN	N	TN
Correlation coefficient	0.1497	0.1497
Mean absolute error	24,2150	24,2150
Root mean squared error	29,3635	29,3635
Relative absolute error	96,9613%	96,9613%
Root relative squared error	99,6671%	99,6671%
Total Number of Instances	268	268

Sources: (Wulandari & Subekti, 2019)

Based on table 5 the results of diabetes positive data using the k-Nearest neighbors (KNN) algorithm with 2 (two) experiments, namely data normalization and without normalization. It is seen that the lowest Root Mean Squared Error (RMSE) value is found in experiments with normalized data and without normalization with a value of 23.33496.

CONCLUSION

Based on the results of research on the prediction of glucose levels in diabetics it can be concluded that the Support Vector Machine Algorithm is better because it has a lower root mean squared error value of 28.9480 compared to comparison algorithms such as Neural Network , Rendom Forest , and K-Nearest Neigbors (KNN). In the plot or visualize error class, it is seen that Instance no. 47 has the highest prediction of the plasma glucose value, 189.2305.

REFERENSI

Al-Khasawneh, A., & Hijazi, H. (2014). A Predictive E-Health Information System: Diagnosing Diabetes Mellitus Using Neural Network Based Decision Support System. *International Journal of Decision Support System Technology*, *6*(4), 31–48. https://doi.org/10.4018/ijdsst.2014100103

Alloghani, M., Aljaaf, A., Hussain, A., Baker, T., Mustafina, J., Al-Jumeily, D., & Khalaf, M. (2019). Implementation of machine learning algorithms to create diabetic patient re-admission profiles. *BMC Medical Informatics and Decision Making*, *19*(Suppl 9), 253. https://doi.org/10.1186/s12911-019-0990-x

Anggraini, A. (2019). *10 Efek Buruk Makanan Manis Pada Tubuh*. Https://Www.Nibble.Id/. https://www.nibble.id/blog/10-efek-buruk-makanan-manis-pada-tubuh/

Belinda, G. (2019). *Kadar Gula Darah Normal dan Cara Mencegah Diabetes*. Honestdocs. https://www.honestdocs.id/kadar-gula-darah-normal

Kahn, M. (1994). *Diabetes Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/diabetes

Kaur, H., & Kumari, V. (2018). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, 1–6. https://doi.org/10.1016/j.aci.2018.12.004

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research. In *Computational and Structural Biotechnology Journal* (Vol. 15, pp. 104–116). Elsevier B.V. https://doi.org/10.1016/j.csbj.2016.12.005

Mukarromah, M., Martha, S., & Ilhamsyah, I. (2015). PERBANDINGAN IMPUTASI MISSING DATA MENGGUNAKAN METODE MEAN DAN METODE ALGORITMA K-MEANS. *BIMASTER*, *4*(3), 305–312. http://jurnal.untan.ac.id/index.php/jbmstr/article/view/12425/

Pappada, S. M., Cameron, B. D., Rosman, P. M., Bourey, R. E., Papadimos, T. J., Olorunto, W., & Borst, M. J. (2011). Neural network-based real-time prediction of glucose in patients with insulin-dependent diabetes. *Diabetes Technology and Therapeutics*, *13*(2), 135–141. https://doi.org/10.1089/dia.2010.0104

Robertson, G., Lehmann, E. D., Sandham, W., & Hamilton, D. (2011). Blood Glucose Prediction Using Artificial Neural Networks Trained with the AIDA Diabetes Simulator: A Proof-of-Concept Pilot Study. *Journal of Electrical and Computer Engineering*, *2011*, 1–11. https://doi.org/doi:10.1155/2011/681786

Sandham, W. A., Lehmann, E. D., Zequera Diaz, M., Hamilton, D. J., Tatti, P., & Walsh, J. (2011). Electrical and computer technology for effective diabetes management and treatment. *Journal of Electrical and Computer*

Engineering, *2011*, 1–3. https://doi.org/10.1155/2011/289359

Saputra, N., Adji, T. B., & Permanasari, A. E. (2016). Analisis sentimen data presiden Jokowi dengan preprocessing normalisasi dan stemming menggunakan metode naive bayes dan SVM. *Jurnal Dinamika Informatika*, *5*(1), 1–12. http://ojs.upy.ac.id/ojs/index.php/dinf/article/view/113

Sarojini, B., Ramaraj, N., & Nickolas, S. (2009). Enhancing the performance of libSVM classifier by kernel f-score feature selection. *Communications in Computer and Information Science*, *40*, 533–543. https://doi.org/10.1007/978-3-642-03547-0_51

Soumya, D., & Srilatha, B. (2011). Late Stage Complications of Diabetes and Insulin Resistance. *Journal of Diabetes & Metabolism*, *02*(09), 1–8. https://doi.org/10.4172/2155-6156.1000167

Sun, S. (2013). A survey of multi-view machine learning. In *Neural Computing and Applications* (Vol. 23, Issues 7–8, pp. 2031–2038). Springer. https://doi.org/10.1007/s00521-013-1362-6

Waila, P., Marisha, S., Singh, V. K., & Singh, M. K. (2012). Evaluating Machine Learning and Unsupervised Semantic Orientation approaches for sentiment analysis of textual reviews. *2012 IEEE International Conference on Computational Intelligence and Computing Research, ICCIC 2012*. https://doi.org/10.1109/ICCIC.2012.6510235

Wulandari, D., & Subekti, A. (2019). *Laporan Akhir Penelitian Mandiri: Prediksi Level Glukosa Pada Penderita Diabetes Dengan Support Vector Regression*.