Journal of Lorem

HTTPS://EJOURNALEJOURNAL.MEDIAKUNKUN.COM/INDEX.PHP/LOREM

LOREM : Computational Engineering and Computer Information Systems
E-ISSN : 3032-3924, P-ISSN : 3032-3894

Comparison of K-Means and K-Medoids Algorithms in Students English Skill Clasterization

Mas'ud Hermansyah[1], Difari Afreyna Fauziah[2], Iqbal Sabilirrasyad[3], M. Faiz Firdausi[4], Abdul Wahid[5]

Information Systems and Technology, Institut Teknologi dan Sains Mandala[1,2]
Software Engineering, Institut Teknologi dan Sains Mandala[3,4,5]

ABSTRACT

Students who have the ability to speak English well can communicate their ideas and ideas in the school environment or with foreigners. English proficiency is not only the ability to speak, but also the ability to understand and produce spoken or written texts which are realized in the four language skills namely listening, speaking, reading and writing. With data mining technology, it is possible to analyze the value of students' English skills. This analysis was carried out by grouping students according to their ability scores in the four skills. In conducting this research, a comparison of the K-Means and K-Medoids clustering methods was used to classify students' English abilities. With the clustering technique, it is hoped that the teacher can adjust the learning model according to the abilities of the students. The purpose of this study is to analyze and process data by comparing the K-Means and K-Medoids methods in clustering English skills scores. Based on the research that has been done, when compared to the K-Means method with K-Medoids, the K-Medoids method is more optimal in terms of the lowest Davies Boldin Index (DBI) value of 0.287 at k=3.

Keywords: *Data Mining, Clustering, K-Means, K-Medoids, Davies Boldin Index*

Corresponding Author:
Mas'ud Hermansyah
- masudhermansyah@gmail.com

Received: September 11, 2023
Revised: October 11, 2023
Accepted: January 02, 2024
Published: January 11, 2024

1. INTRODUCTION

For high school students, English is a compulsory subject that is taught in developing students' knowledge, language skills, and a positive attitude towards English. So that the English given is presented in an interesting, quality and in accordance with existing developments. rites for high school students so that these students are able to compete in the field of science and are able to compete with other countries. Students who have the ability to speak English well can communicate their ideas and ideas in the school environment or with foreigners. However, there are still many students at the senior high school level who still experience difficulties in conveying ideas, thoughts and questions in English using good and correct spoken language (Tambusai & Nasution, 2022).

Students find it difficult to speak English because English is not used as everyday language. Language difficulties are a real form of proficiency and the ability to listen, speak, read, write and reason (Farid et al., 2022). Learning difficulties are divided into two groups, namely, difficulties related to the development of developmental learning disabilities including motor, perceptual, language learning difficulties, communication and learning difficulties in adjusting social behavior. While the second difficulty is related to academic values (academic learning disabilities), namely the failure to achieve achievements that are not in accordance with the expected capacity (Silalahi et al., 2022).

English is often a problem for students. According to (Silalahi et al., 2022) in his research explained that the first reason that was most often stated was because English was not the mother tongue so it was difficult to pronounce it. The second reason is feeling lazy to practice listening, speaking, reading and writing so that it makes English even more difficult to understand. This second reason should be a provision for teaching English in class. However, some educators often forget to present the English language needs according to the needs of their students. The aim of this subject is to equip students with active communication skills in English, namely the ability to listen, read and write. Mastery of English is also a means to boost Indonesia's human resources, which according to the Human Development Index are in the lowest category in Asia. Global competition in all English that demands an increase in the quality of human resources, including teaching staff, as the spearhead. The school's output must really be of good quality in order to be competitive and have a high bargaining position. One of the efforts to realize the above is to improve the quality of learning English. Mastery of English will open their horizons to the development of science and technology, including education which is currently easily accessible from various sources.

With the existence of data mining technology, an analysis of students' English skills can be carried out. This analysis was carried out by grouping students according to their listening, speaking, reading and writing abilities. The division of study groups using the clustering process is done by dividing a group of students into subsets called clusters. Clustering is a method in data mining that is useful in analyzing data to make it more accurate in solving data grouping problems or dividing a data set into several subsets, Like machine learning models is used to address more complex patterns and interactions among variables (Wiranto, Sabilirrasyad, et al., 2023). The purpose of clustering is to assign data into a group so that the relationship between members in the same cluster becomes strong, while the relationship between members in different clusters becomes weak. (Agustina et al., 2012). Objects in a cluster have similar characteristics but have different characteristics from objects in other clusters. Therefore, clustering is very useful in assigning unknown groups or clusters to the data (Dacwanda & Nataliani, 2021).

In conducting this research, a comparison of the K-Means and K-Medoids clustering methods was used to classify students' English abilities. The use of a comparison of the two methods is to find the best group with a comparison of the Davies Bouldin Index (DBI) values (Hermansyah et al., 2023). With the data mining clustering technique, it is expected that schools and parties in the education sector can record students with their respective abilities and can be taught using the right method so as to improve the quality of students' English and student academic achievement, such as increases harvest success, and ensures better quality output (Wiranto, Rohim, et al., 2023). The focus of this research is on classifying student achievement at Lab Business School Tangerang Vocational High School with a comparison of the K-Means and K-Medoids clustering methods, where students are grouped according to their level of ability in listening, speaking, reading and writing.

2. METHODS

2.1 Data Mining

Data mining as a process of finding useful information from a large database warehouse. Data mining can also be interpreted as extracting information from large data sets to assist in decision making. Data mining or knowledge discovery in database is the process of resourcing and using data to find patterns or relationships from large data sets. the results of the data mining process can be used as an evaluation of future decision making. The public definition of data mining is a method of searching for previously unknown hidden knowledge patterns from a very large set of data in databases, data warehouses, or other storage media. Data mining is used to explore added value in the form of information that is not known manually from a database. Information is obtained by extracting and recognizing important or interesting patterns from the data contained in the database (P. N. Harahap & Sulindawaty, 2019).

2.2 Clustering

Clustering on a data is a step to classify data sets whose class attributes have not been described, conceptually clustering is to maximize and minimize intra-class similarities. For example, there is a set of objects, the first process can be clustered into several sets of classes, then it becomes an ordered set so that it can be derived based on certain classification groups. Cluster can also be interpreted as a group. then the clustering analysis will basically produce a number of clusters (groups). Before we do the analysis, it is necessary to apply the understanding that a set of certain data already has similarities among its members. Therefore, each member who has similar characteristics is grouped into one or more of a group. The purpose of data clustering is to minimize the objective function set in the clustering process, and generally always minimize the variation of a cluster and maximize the variation between clusters. (Muliono & Sembiring, 2019).

2.3 K-Means Clustering

K-Means clustering is a method that belongs to non-hierarchical clustering where every object included in the group is the same and correlated objects. Data belonging to groups has a greater degree of similarity and also has a greater degree of difference from other groups (B. Harahap, 2019).

The K-Means algorithm is an algorithm that works by partitioning data into clusters, so that data that is similar is in the same cluster and data that is dissimilar is in another cluster. (Rohmawati et al., 2015).

The following are the steps contained in the K-Means algorithm:

1. Determine the number of clusters (k), set the cluster center randomly.
2. Calculate the distance of each data to the center of the cluster.
3. Group data into clusters with the shortest distance.
4. Compute the new cluster center.
5. Repeat steps 2 (two) to 4 (four) so that no more data is moved to another cluster.

The clustering process begins with identifying clustered data, using the Euclidean Distance formula as shown in equation (1).

$$d_{ij} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2} \tag{1}$$

Keterangan:

D (i,j) = The distance between i to the data center cluster j
X ki = The i data on the k data attribute
X kj = The jth center point on the kth attribute

$$C = \frac{\sum m}{n} \tag{2}$$

Equation 2 explains where C is a data centroid, m is a data member belonging to a certain centroid and n is the number of data members belonging to a certain centroid.

2.4 K-Medoids Clustering

The K-Medoids or Patition Around Medoids (PAM) algorithm was developed by Leonard Kaufman and Peter J. Rousseeuw in 1987. The PAM algorithm includes the Partitioning clustering method for grouping a group of objects into clusters. Medoid is a cluster representation in PAM of a set of objects that represent clusters (Wira et al., 2019). K-Medoids is a clustering algorithm that is almost the same as the K-Means algorithm. The difference is that K-Medoids uses an object as the cluster center for each cluster, while K-Means uses the average value as the cluster center for each cluster. The K-Medoids algorithm is used to overcome the weaknesses of the K-Means algorithm which is very sensitive to noise and outliers, where objects with large values allow deviations in the data distribution. (Febrianti et al., 2019). There are several steps in the K-Medoids algorithm namely (Rahmah et al., 2022):

1. Initialize k cluster centers (number of clusters).
2. Count each object to the closest cluster using the Euclidian Distance measure equation:

$$d(x,y) = \sqrt{\sum_{i-1}^{n}(xi - yi)^2} \quad (3)$$

Keterangan:
d(x,y) Information: = distance between data I and data j
xi1 = value of the first attribute from the i-th data
yj1 = value of the first attribute of the jth
n = the number of attributes used.

1. Randomly select objects as points that are not medoids.
2. Calculate the object distance in each cluster with non-medoids candidates.
3. Calculate the total deviation (S), if the new TD<old TD, change the position of the new medoid, then it becomes a new medoid.
4. Perform steps 3-5 so that the medoid does not change.

2.5 Davies Bouldin Index Davies (DBI)

Davies Bouldin Index (DBI) is calculating the average value of each point in the data set. Calculation of the value of each point is the sum of the compactness values divided by the distance between the two cluster center points as separation (Hermansyah et al., 2020).

In the process of assessing the resulting model, the Davies Bouldin Index is used. DBI is used to optimize the distance outside the cluster and minimize the distance inside the cluster which can be calculated by the following equation 4 (Mahartika & Wibowo, 2019):

$$S_i = \frac{1}{|ci|}\sum_{x\in ci}\{|x - z_i|\} \quad (4)$$

Where *ci* is the number of points contained in cluster i, x is data, and *zi* is the centroid of cluster i. While the distance between clusters is defined in Equation 5 below:

$$d_{ij} = \left|Z_i - Z_j\right| \quad (5)$$

Where z_i is the centroid of cluster i and z_j is the centroid of cluster j. Calculation of the distance d_{ij} can use euclidean distance. Then define $R_{i,qt}$ for the c_i cluster in Equation 6 below:

$$R_{i,qt} = \underset{j,j\neq 1}{max}\{\frac{S_{i,i}+S_{j,q}}{d_{ij,t}}\} \quad (6)$$

Furthermore, the Davies Bouldin Index is defined in Equation 7 below:

$$DBI = \frac{1}{k}\sum_{i=1}^{k} Ri, qt \quad (7)$$

From this equation, k is the number of clusters used. With the condition that the smaller the value obtained from the DBI calculation (non-negative >= 0), the better the clusters obtained from cluster grouping using the K-means method used.

3. RESULTS AND DISCUSSION

The data source was taken from the academic scores of the English subject skills of class IX students of the Multimedia Expertise Program, which consisted of 76 students. Class XI student data that is processed is data for 2022 and has the following attributes:

1. Absence number
2. Name of Student
3. Gender
4. The value of listening
5. The value of speaking
6. The value of reading
7. The value of writing

Of the several attributes previously mentioned, only four attributes were selected to be used in the clustering process. that is:

1. The value of listening
2. The value of speaking
3. The value of reading
4. The value of writing

Of the 4 attributes used in the calculation process, namely the ability to value listening, speaking, reading and writing. Cleansing data is to reduce noise that can affect calculations. in the data cleansing process, data that has vacant values on 4 attributes are not used, so the data used is 72 students out of 76 students. data that has been processed data cleansing can be seen in table 1.

Table 1. Data Cleansing Results

Absence number	Listening	Speaking	Reading	Writing
1.	86	85	88	82
2.	79	72	76	74
3.	68	72	72	78
4.	84	83	85	86
5.	80	73	80	76
…	…	…	…	…
71.	74	78	74	79
72.	70	75	80	68

Source: data processing, 2023

This study aims to compare and find the best cluster distribution by measuring the Davies Bouldin Index Davies (DBI) scores to classify students' abilities in English proficiency. Furthermore, the mining process is carried out with the aim of finding information or patterns for clustering using a comparison of the K-Means and K-Medoids algorithms. The implementation of the K-Means and K-Medoids algorithms in this study used the Rapidminer software. Implementation using the Rapidminer software can be seen in Figure 1.

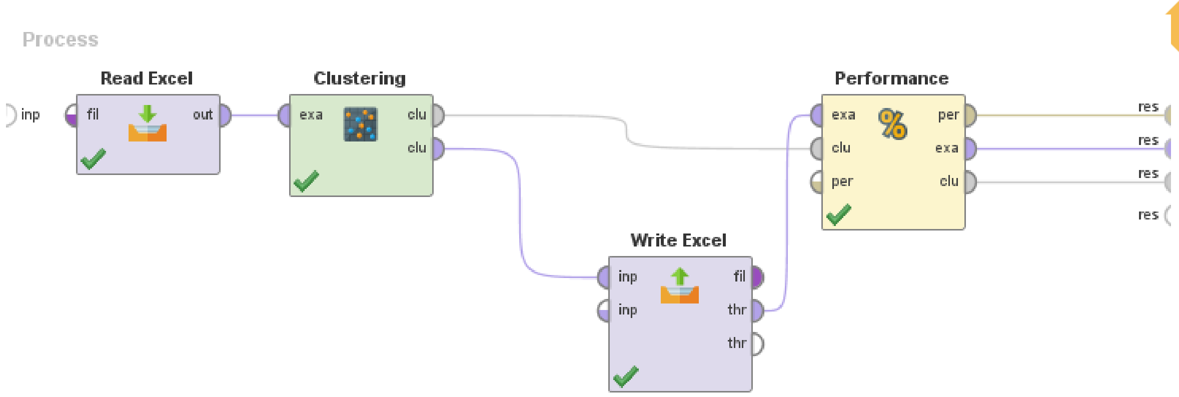

Figure 1. Modeling K-Means and K-Medoids Clustering with Rapidminer
Source: data processing, 2023

1. Read excel is an operator to read the example set from the specified excel file.
2. Clustering is an operator that performs a grouping process using the K-Means and K-Medoids algorithms.
3. Write excel is an operator for creating clustering results reports with type file.xlsx.
4. Performance is the operator used to evaluate the performance of the clustering method based on the centroid.

The results of the K-Means and K-Medoids comparison in grouping students' English language ability data by testing 5 times in clusters 2 to 6 are shown in table 2 (K-Means) and table 3 (K-Medois).

Table 2. K-Means Modeling Results

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Total data k = 2	57	15	-	-	-	-
Total data k = 3	24	35	13	-	-	-
Total data k = 4	13	16	19	24	-	-
Total data k = 5	5	9	24	8	16	-
Total data k = 6	12	7	20	13	13	7

Source: data processing, 2023

Table 3. K-Medoids Modeling Results

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Total data k = 2	40	32	-	-	-	-
Total data k = 3	33	17	22	-	-	-
Total data k = 4	26	18	12	16	-	-
Total data k = 5	21	12	12	16	11	-
Total data k = 6	12	18	23	7	5	7

Source: data processing, 2023

After clustering using the K-Means and K-Medoids methods, the next stage of this research is to determine the most optimal number of clusters using rapidminer software as seen from the Davies Bouldin Index (DBI). This process is carried out to determine the DBI value of each method in each cluster. The test was carried out from clusters k=2 to k=6. The following results of a comparison of DBI can be seen in Figure 2.

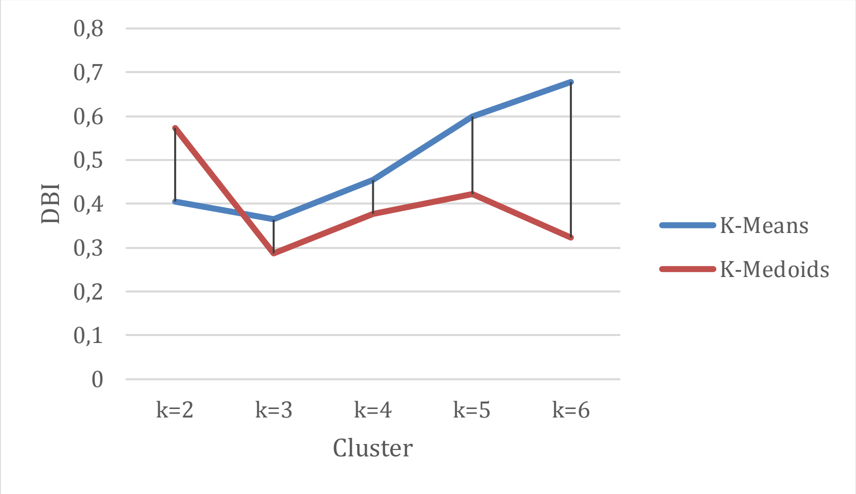

Figure 2. Comparison of DBI K-Means and K-Medoids Algorithms
Source: data processing, 2023

To see the results of testing each cluster from the Rapidminer software on the K-Means and K-Medoids methods, see Table 4.

Table 4. Comparison Results of the Davies Bouldin Index (DBI)

Cluster	**Method**	
	K-Means	**K-Medoids**
k-2	0,404	0,573
k-3	0,365	0,287
k-4	0,454	0,377
k-5	0,599	0,423
k-6	0,678	0,323

Source: data processing, 2023

From the results of the experiments conducted, the K-Medoids algorithm with Rapidminer, the number of 3 clusters produces better cluster quality compared to the number of clusters 2, 4, 5, and 6. The results of the cluster evaluation show that the K-Means algorithm with the number of clusters 3 is more optimal with the smallest DBI value, of 0.365. Whereas the K-Medoids algorithm with rapidminer, the number of 3 clusters also produces better cluster quality compared to the number of clusters 2, 4, 5, and 6. The results of the cluster evaluation show that the K-Medoids algorithm with the number of clusters 3 is more optimal with the highest DBI value. small, amounting to 0.287.

Based on table 4, the results of this comparison get an analysis that the score of English skills with the 3 best clusters classifies the smart cluster as many as 33 students, the medium cluster is 17 students, and as many as 22 students are sufficient. The results of the cluster grouping can be used as a reference and consideration for grouping study groups.

4. CONCLUSION

After analyzing and processing the data by comparing the K-Means and K-Medoids methods in clustering English skill scores, it can be concluded that the DBI values obtained from the K-Means and K-Medoid methods with an experiment of forming six clusters produce the smallest DBI values in the K-Medoids method is the value of k = 3, namely 0.287. As a performance comparison, cluster formation using the K-Means method also has the smallest DBI value at k=3, which is equal to 0.365. Thus, the most optimal formation of clusters in the clustering of English skill scores is to use the K-Medoids method. The analysis shows that the score of English skills with the best 3 clusters classifies the smart cluster as many as 33 students, the medium cluster as many as 17 students, and as many as 22 students as sufficient. The results of the cluster grouping can be used as a reference and consideration for grouping study groups.

REFERENCES

Agustina, S., Yhudo, D., Santoso, H., Marnasusanto, N., Tirtana, A., & Khusnu, F. (2012). Clustering Kualitas Beras Berdasarkan Ciri Fisik Menggunakan Metode K-Means. *Clustering K-Means*, 1–7.

Dacwanda, D. O., & Nataliani, Y. (2021). Implementasi k-Means Clustering untuk Analisis Nilai Akademik Siswa Berdasarkan Nilai Pengetahuan dan Keterampilan. *Aiti : Jurnal Teknologi Informasi*, *18*(2), 125–138. https://doi.org/10.24246/aiti.v18i2.125-138

Farid, M., Tanasal, A. A., Puli, A. R. F. A., Alamin, R. L., & Renaldi. (2022). Program English Area : Upaya Meningkatkan Kemampuan Berbahasa Inggris Siswa SMK. *Jurnal Lepa-Lepa Open*, *2*(5), 1291–1299.

Febrianti, E., Sembiring, R. W., & Suhada, D. (2019). Mengkluster Jumlah Kabupaten/Kota Yang Melaksanakan Kawasan Tanpa Rokok (KTR) Di 50% Sekolah Menurut Provinsi Dengan K-Medoids. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *3*(1), 637–644. https://doi.org/10.30865/komik.v3i1.1672

Harahap, B. (2019). Penerapan Algoritma K-Means Untuk Menentukan Bahan Bangunan Laris (Studi Kasus Pada UD. Toko Bangunan YD Indarung). *Regional Development Industry & Health Science, Technology and Art of Life*, 394–403.

Harahap, P. N., & Sulindawaty, S. (2019). Implementasi Data Mining Dalam Memprediksi Transaksi Penjualan Menggunakan Algoritma Apriori (Studi Kasus PT.Arma Anugerah Abadi Cabang Sei Rampah). *MATICS: Jurnal Ilmu Komputer Dan Teknologi Informasi*, *11*(2), 46–50. https://doi.org/10.18860/mat.v11i2.7821

Hermansyah, M., Hamdan, R. A., Sidik, F., & Wibowo, A. (2020). Klasterisasi Data Travel Umroh di Marketplace Umroh.com Menggunakan Metode K-Means. *Jurnal Ilmu Komputer*, *13*(2), 8. https://doi.org/10.24843/jik.2020.v13.i02.p06

Hermansyah, M., Prasetyo, N. A., Ansori, Y., FIrdausi, M. F., & Wahid, A. (2023). Implementation of K-Means Clustering for Analysis Students English Proficiency. *Journal of Education Science and Technology*, *1*(6), 31–35.

Mahartika, I. R., & Wibowo, A. (2019). Data Mining Klasterisasi dengan Algoritme K-Means untuk Pengelompokkan Provinsi Berdasarkan Konsumsi Bahan Bakar Minyak Nasional. *Prosiding Seminar Nasional Sisfotek*.

Muliono, R., & Sembiring, Z. (2019). Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen. *CESS (Journal of Computer Engineering, System and Science)*, *4*(2), 2502–2714.

Rahmah, E., Haerani, E., Nazir, A., & Ramadhani, S. (2022). Penerapan Algoritma K-Medoids Clustering Untuk Menentukan Srategi Promosi Pada Data Mahasiswa (Studi Kasus : Stikes Perintis Padang). *Jurnal Nasional Komputasi Dan Teknologi Informasi (JNKTI)*, *5*(3), 556–564. https://doi.org/10.32672/jnkti.v5i3.4355

Rohmawati, N., Defiyanti, S., & Jajuli, M. (2015). Implementasi Algoritma K-Means Dalam Pengklasteran Mahasiswa Pelamar Beasiswa. *Jitter : Jurnal Ilmiah Teknologi Informasi*

Terapan, *1*(2), 62–68.

Silalahi, M., Purba, A., Benarita, B., Matondang, M. K. ., Sipayung, R. W., Silalahi, T. F., Saragih, N., Girsang, S. E., Damanik, I. J., & Sibuea, B. (2022). Analisis Kesulitan Belajar Bahasa Inggris Siswa Sma Negeri 1 Narumonda Kabupaten Tobasa. *Community Development Journal : Jurnal Pengabdian Masyarakat*, *3*(2), 728–732. https://doi.org/10.31004/cdj.v3i2.4686

Tambusai, A., & Nasution, K. (2022). Tingkat Pemahaman Bahasa Inggris Bagi Siswa Sekolah Menegah Atas (SMA). *Jurnal Pema Tarbiyah*, *1*(1), 44–53.

Wira, B., Budianto, A. E., & Wiguna, A. S. (2019). Implementasi Metode K-Medoids Clustering untuk Mengetahui Pola Pemilihan Program Studi Mahasiswa Baru Tahun 2018 di Universitas Kanjuruhan Malang. *Jurnal Terapan Sains & Teknologi*, *1*(3), 54–69.

Wiranto, F., Rohim, M. A., Firdausi, M. F., Muliawan, A., & Afrianto, E. (2023). *Optimizing Coffee Crop Selection for Plant-Ready Coffee Fields : A Study Using the " Predict . in " Decision Support System at the Coffee and Cocoa Research Center in Jember*. *6*(1), 42–47.

Wiranto, F., Sabilirrasyad, I., Hermansyah, M., Mandala, S., Wiranto, F., & Mandala, S. (2023). *Optimizing Forecasting of Dow Jones Stock Index in New York amid Uncertain Global Conditions in 2023 : A Combined Approach of ARIMA and Machine Learning Models*. *1*, 73–88.