Journal of Computer Science, Information Technology and Telecommunication Engineering (JCoSITTE)
Vol. 5, No. 1, March 2024, pp. 504~509
ISSN: 2721-3838, DOI: 10.30596/jcositte.v5i1.17991

Application of Data Mining to Predict Birth Rates in Medan City Using the K-Nearest Neighbor Method

Alma Irawanti Putri[1], Mhd. Furqan[2], Suhardi[3]
[1,2,3]Department of Computer Science, Faculty of Science and Technology, Universitas Islam Negeri Sumatera Utara Medan

ABSTRACT

The birth rate of babies in Indonesia tends to increase every month, based on this fact, the population in Indonesia is increasing over time. One of the contributing factors is increasingly sophisticated technology, so that a country's birth rate can be accelerated, and if this event occurs continuously it will have an impact on population density which will occur not only in Indonesia, but also throughouty the world. Therefore, birth rate predictions are needed for planning and public policy in the fields of health and social welfare. One of them is using data mining techniques to predict the number of births in Medan City using the KNN method. KNN is a classification method based on the neighborhood value between training data and test data. The prediction results will be compared with actual data to measure the accuracy of predictions on birth data totaling 131 data. The accuracy results obtained were 83.9% with a total of 4,413 births and 8,485 pregnant women

Keyword: Birthrate; Data Mining; K-Nearest Neighbor.

Corresponding Author:
Alma Irawani Putri,
Department of Computer Science
Universitas Islam Negeri Sumatera Utara
Jl. Lapangan Golf No.120, Medan Tuntungan, 20138, Indonesia.
Email: irawani668@gmail.com

Article history:
Received : Dec 29th 2023
Revised : Jan 4th 2023
Accepted : Jan 30th 2024

1. INTRODUCTION

The birth rate in Indonesia tends to increase every month, based on this fact the population in Indonesia is increasing over time (Novrizaldi, 2021). One of the causal factors is increasingly sophisticated technology, so that a country's birth rate can be accelerated, and if this event continues to occur continuously it will have an impact on population density which will occur not only in Indonesia, but also throughout the world. (Syahra et al., 2019). High birth rates require special attention in handling by the government. The attention and handling that the government can do is to create a comprehensive Family Planning (KB) program by urging the public to limit births to 2 children per family. (Parihah et al., 2020). This can be done by analyzing data to find out predictions of the increase in the number of births each year using data mining techniques. (Idris, 2020).

Data mining is data that is used to analyze large data by finding clear relationships and concluding things that were not previously known. (Pratama et al., 2022). In this research, data mining was carried out by predicting birth rates using the K-Nearest Neighbor method (KNN). KNN is a supervised learning algorithm that makes an approach based on the closest distance or neighborhood value. (Indahsari & Kurniawan, 2019). KNN is also a prediction method based on the majority of class labels by classifying new data based on attributes and training data (Prasetyaningrum, 2021).

K-Nearest Neighbor (KNN) is an approach to calculating the closeness between new cases and old cases based on matching the weights of a number of existing features. (Syukri Mustafa & Wayan Simpen, 2019). The process carried out by KNN is as follows (Sekar Seruni et al., 2020):

1. Determine the neighborhood value K
2. Find the Euclidean distance value for each test data to the training data using the following equation:
   $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
3. Rank the Euclidean distance results from smallest to largest.
4. Collect nearest neighbor classification (category Y).
5. Find the prediction results for the query instance value based on the majority category.

Research that is relevant to this research is research on birth age prediction using the K-Nearest Neighbor method, where this research has an accuracy rate of 96% using 560 birth data. (Indahsari & Kurniawan, 2019). Then there is research on determining bad credit for electronic goods using data mining techniques with the K-Nearest Neighbor algorithm. This research aims to stabilize the company's finances by making predictions in determining customers who have the potential to carry bad credit which have been previously recorded. The results of using this method are accurate. which is good and correct (Silvilestari, 2021). Meanwhile, weather prediction research in Indonesia using KNN has an accuracy of 89% with a total of 2.898 training data and 725 test data. (Rangkuti et al., 2021).

From the results of the analysis of relevant research, further research can be carried out in the form of birth rate predictions using data mining with the K-Nearest Neighbor method. In this research, success or accuracy testing will be carried out using the confusion matrix method. The equations in the confusion matrix are as follows (Sakti et al., 2018):

$$Akurasi = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100\%$$

$$Presisi = \frac{\text{TP}}{\text{TP+FP}} \times 100\%$$

$$Recall = \frac{\text{TP}}{\text{TP+FN}} \times 100\%$$

$$F1\ Score = \frac{2 \times (\text{Recall} \times \text{Presisi})}{\text{Recall+ Presisi}}$$

Where:
TP (True-Positive) = The number of actual positive class data and positive prediction results
TN (True-Negative) = The number of actual negative class data and negative prediction results
FP (False-Positive) = The number of actual negative class data and positive prediction results
FN (False-Negative) = The number of actual positive class data and negative prediction results

2. RESEARCH METHOD

This research uses data from the Central Statistics Agency on the website: https://bps.go.id/ from 2010 to 2020, totaling 131 data based on sub-district area data with attributes such as: Month/Year, Number of Births, and Number of Mothers Pregnant. The process that will be carried out is to normalize the data first using the following min-max normalization equation (Ryan et al., 2023):

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where:
X = Original value
$X_{normalized}$ = Normalization results of X
X_{min} = Lowest value in the data range
X_{max} = The highest value in the data range

After normalizing the data, the next process can be carried out, namely data transformation by creating data groups based on "Low" and "High" classes by calculating the average of the number of birth rates and the number of pregnant women first in order to make it easier to predict birth rates. Where if the number of births and pregnant women is smaller than the average then it is in the "Low" class group, conversely if it is higher or the same as the average value then it is in the "High" class group. This is done to create classes in the training data.

The next process predicts birth rates on test data against training data using K-Nearest Neighbor with a neighbor value (K=5). Then the process of testing the accuracy of the KNN test results using a confusion matrix will be carried out. This process can also be seen in Figure 1 below:

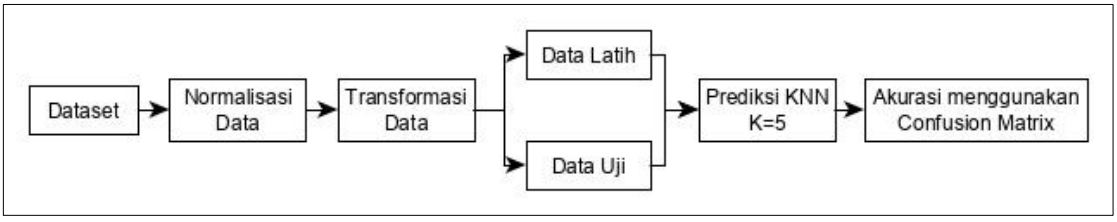

Fig 1. Birth rate prediction framework using KNN

3. RESULTS AND DISCUSSION

This stage was tested for birth rate predictions from Central Statistics Agency data using K-Nearest Neighbor and testing the accuracy of KNN prediction results using a confusion matrix. The results of birth rate predictions using KNN are as shown in the following table:

Table 1. Test data

Number of Births (Soul)	Number of Pregnant Women (Soul)	Euclidean Distance	Class
Dec, 2020	8.485	4.413	?

Table 2. Prediction results using KNN with K=5

Ranking Data	Number of Births (Soul)	Number of Pregnant Women (Soul)	Ecludian Distance	Class
1	8.201	4.257	324.02	High
2	8.045	4.012	595.32	High
3	8.850	5.017	705.72	High
4	7.806	4.029	780.06	High
5	7.712	4.588	792.56	High
6	7.663	4.501	826.70	High
7	8.413	3.514	901.88	High
8	8.869	3.425	1,060.00	High
9	7.549	3.791	1,123.82	High
10	7.372	4.236	1,126.99	High
11	7.349	4.471	1,137.48	High
12	7.450	3.771	1,217.94	High
13	7.758	3.239	1,380.87	High
14	7.638	3.256	1,433.90	High
15	7.093	3.570	1,627.36	High
16	7.125	3.518	1,628.07	High
17	6.927	5.101	1,703.15	High
18	6.833	4.889	1,719.21	High
19	7.380	3.041	1,761.65	High
20	7.192	3.101	1,842.06	High
21	6.814	3.603	1,856.97	High
22	7.145	3.025	1,929.29	High
23	7.121	2.991	1,970.43	High
24	6.691	3.577	1,979.23	High
25	7.187	2.858	2,025.54	High
26	7.013	2.994	2,044.59	High
27	6.476	3.869	2,081.35	High
28	6.510	3.428	2,207.00	High
29	6.753	3.014	2,226.44	High
30	6.266	4.135	2,236.35	High
31	6.352	3.705	2,247.43	High
32	6.312	5.130	2,288.23	High
33	6.185	4.216	2,308.42	High
34	6.250	3.610	2,374.88	High
35	6.169	4.974	2,382.98	High
36	6.817	6.125	2,390.22	High
37	6.087	4.305	2,400.43	High
38	6.110	4.762	2,400.51	High
39	6.047	4.457	2,438.40	High
40	7.054	2.413	2,459.22	High
41	6.214	3.319	2,520.77	High
42	6.028	5.147	2,564.29	High
43	7.388	6.783	2,611.57	High
44	5.890	4.708	2,611.71	High
45	6.271	6.008	2,728.70	High
46	5.789	3.915	2,741.61	High
47	5.858	3.371	2,826.11	High
48	5.728	3.625	2,867.40	High
49	5.715	3.647	2,873.96	High
50	5.694	5.125	2,880.39	High
51	5.748	3.234	2,980.14	High
52	5.415	4.328	3,071.18	High

53	5.433	3.728	3,127.93	High
54	5.471	3.509	3,146.65	High
55	5.310	4.021	3,199.11	High
56	5.505	3.217	3,211.05	High
57	7.124	1.400	3,306.13	High
58	5.851	6.507	3,364.94	High
59	5.158	3.899	3,366.47	High
60	5.104	3.949	3,412.69	High
61	5.135	3.667	3,432.06	High
62	5.070	4.029	3,436.52	Low
63	5.010	3.758	3,536.19	Low
64	4.941	3.998	3,568.22	Low
65	5.021	3.412	3,605.73	Low
66	5.103	5.729	3,629.02	High
67	5.053	3.203	3,639.06	Low
68	5.166	2.845	3,670.75	High
69	4.805	4.655	3,687.95	Low
70	9.518	8.650	3,695.32	High
71	4.923	3.029	3,821.43	Low
72	4.618	5.126	3,932.18	Low
73	4.523	4035	3,979.99	Low
74	6.204	7.732	4,027.25	High
75	4.319	4.815	4,185.35	Low
76	4.270	4.405	4,215.01	Low
77	4.236	3.918	4,277.74	Low
78	4.269	3.517	4,310.16	Low
79	4.153	4118	4,342.03	Low
80	4.103	5.031	4,425.36	Low
81	4.227	3.039	4,474.20	Low
82	3.968	3.981	4,537.61	Low
83	3.950	4.597	4,538.73	Low
84	3.901	4.901	4,609.90	Low
85	3.790	4.479	4,695.46	Low
86	3.830	5.209	4,722.57	Low
87	3.759	4.832	4,744.54	Low
88	5.647	4.200	4,898.81	High
89	3.721	3.250	4,903.90	Low
90	3.613	3.813	4,908.81	Low
91	12.861	2.1150	4,942.69	High
92	3.546	4.038	4,953.22	Low
93	10.026	9.275	5,100.37	High
94	3.350	4.224	5,138.48	Low
95	3.282	4.908	5,226.49	Low
96	3.242	4.105	5,252.04	Low
97	3.195	4.231	5,293.13	Low
98	12.594	1.052	5,308.50	High
99	3.179	3928	5,328.12	Low
100	3.115	4.605	5,373.43	Low
101	3.261	3.027	5,404.74	Low
102	3.305	2.779	5,431.61	Low
103	3.026	4.689	5,465.97	Low
104	3.051	3804	5,468.02	Low
105	2.941	4364	5,544.22	Low
106	3.003	3.395	5,575.72	Low
107	2.872	4.431	5,613.03	Low
108	3.055	2.968	5,618.98	Low
109	2.813	4.770	5,683.22	Low
110	2.653	4.328	5,832.62	Low
111	2.614	4.367	5,871.18	Low
112	2.579	4.251	5,908.22	Low
113	2.554	4.932	5,953.66	Low
114	2.548	3.724	5,976.85	Low
115	2479	5.625	6,127.07	Low
116	2.017	4217	6,470.97	Low
117	2.194	2.750	6,507.09	Low
118	1.883	4.739	6,610.04	Low
119	4.715	9.866	6,629.34	Low
120	3.029	4.210	6,760.47	Low
121	1479	5.233	7,053.82	Low

122	1250	5.713	7,350.87	Low
123	1168	5.517	7,399.82	Low
124	1101	5.012	7,408.26	Low
125	1076	3.907	7,426.26	Low
126	895	2.973	7,725.39	Low
127	802	3.015	7,809.15	Low
128	663	4.713	7,827.75	Low
129	543	4.293	7,942.91	Low
130	741	2.572	7,959.83	Low
131	380	4.571	8,106.54	Low

Based on the results of testing test data against training data using KNN with a neighborhood value (K=5), the predicted birth rate in December 2020 is in the "High" class. The accuracy results of the confusion matrix from predicting birth rates using KNN are as follows:

Table 3. Confusion Matrix

Predicted/Actual	True	False
Positive	80	12
Negative	9	30

From table 3 it can be seen that the accuracy is 83,9%, precision is 86,9%, recall is 72,7% and f1-score is 79,16.

4. CONCLUSION

Based on the process of applying the KNN method in predicting birth rates, it can be concluded that the KNN method with a neighborhood value (K=5) can predict birth rates with a truth or accuracy of 83,9% with a total of 4.413 births and a total of 8.485 pregnant women. from the total training data 131 data.

REFERENCES

Idris, M. (2020). Implementasi-Data Mining-Dengan AlgoritmaaNaïve Bayess Untuk Memprediksi TKP-Kriminalitas Di KabupatennPonorogo. *Paper Knowledge . Toward a Media History of Documents*, *7*(1), 1–33.

Indahsari, D. K., & Kurniawan, Y. I. (2019). AplikasiiPrediksi Usia KelahirannDengan Metode K-NearestNeighbor. *Jurnal Kebidanan*, *11*(01), 1. https://doi.org/10.35872/jurkeb.v11i01.335

Novrizaldi. (2021). *Hasil Surveii Penduduk 2020 Peluang Indonesia Maksimalkann-Bonus Demografi*. KEMENKO-PMK. https://www.kemenkopmk.go.id/hasil-survei-penduduk-2020-peluang-indonesia-maksimalkan-bonus-demografi

Parihah, N. I., Hartini, S., & Siregar, J. (2020). Prediksi_Angka KelahirannBayi Pada Desa ATridaya Saktii Dengan Menggunakan Algoritmaa NaiveBayes. *Journal of Students' Research in Computer Science*, *1*(2), 77–88. https://doi.org/10.31599/jsrcs.v1i2.423

Prasetyaningrum, A. U. H. dan P. W. (2021). Penerapan Dataa Minning untuk Prediksi Layanan Produksi Indihome Menggunakan Metode K-NearestNeighbor. *JURNAL INFORMATION SYSTEM & ARTIFICIAL INTELLIGENCE*, *1*(2),100–107.http://repository.uin-suska.ac.id/23008/%0Ahttp://repository.uin-suska.ac.id/23008/2/NURFITRIANTI.pdf

Pratama, F. D., Zufria, I., & Triase, T. (2022). Implementasi Data Minning Menggunakan Algoritmaa NaïveBayes Untuk Klasifikassi Penerima Program Indonesia Pintar. *Rabit : Jurnal Teknologi Dan Sistem Informasi Univrab*, *7*(1), 77–84. https://doi.org/10.36341/rabit.v7i1.2217

Rangkuti, M. Y. R., Alfansyuri, M. V., & Gunawan, W. (2021). PenerapanAlgoritma K-NearestNeighbor (Knn) Dalam MemprediksiDan MenghitungTingkat AkurasiData CuacaDi Indonesia. *Hexagon Jurnal Teknik Dan Sains*, *2*(2), 11–16. https://doi.org/10.36761/hexagon.v2i2.1082

Ryan, I. M., Dhita, P., Ayu, G., & Mastrika, V. (2023). ImplementasiAlgoritma KNNuntukMemprediksi PerformaSiswa Sekolah. *Jurnal Nasional Teknologi Informasi Da Aplikasinya*, *1*(3), 819–826.

Sakti, O., Prakasa, Y., Lhaksamana, K. M., Informatika, F., Telkom, U., Mining, T., Classifier, K. N., & Distance, E. (2018). *KlasifikasiTekssDengan Menggunakan Algorittma K-Nearest*. *5*(3), 8237–8248.

Sekar Seruni, D., Tanzil Furqon, M., & Cahya Wihandika, R. (2020). Siistem Prediksi PertumbuhanJumlah Penduduk Kota Malang MenggunakanMetode K-NearestNeighborRegression. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *4*(4), 1075–1082.

Silvilestari, S. (2021). Data Minning Menggunakan AlgoritmaaK-NearestNeighbor Dalam Menentukan KreditMacet BarangElektronik. *Jurnal Media Informatika Budidarma*, *5*(3), 1063. https://doi.org/10.30865/mib.v5i3.3100

Syahra, Y., Santoso, I., & Kustini, R. (2019). Implementassi Data Minning Untuk PrediksiAngka Kelahirran BayiPada Desa SibolangitMenggunakan Multi Regresi. *Seminar Nasional Sains & Teknologi Informasi (SENSASI)*, *1*, 687–

690.
Syukri Mustafa, M., & Wayan Simpen, I. (2019). Implementassi AlgoritmaaK-Nearest Neighbor (KNN) Untuk Mempredikksi Pasien Terkenna Penyakit Diabettes Pada Puskesmas Manyampa Kabupaten Bulukumba. *Februari*, *2019*(1), 1–10.