International Journal of Computer, Network Security and Information System (IJCONSIST)
Vol. 5 No. 1, September 2023

Prediction of Birth Rates Using the Naive Bayes Algorithm in the North Sumatra Region

Dhian Satria Yudha Kartika
Department of Digital Business, Faculty of Computer Science, UPN Veteran Jawa Timur
Surabaya, Jawa Timur
dhian.satria@upnjatim.ac.id

Brillyan Putra
Department of Information System, Faculty of Computer Science, UPN Veteran Jawa Timur
Surabaya, Jawa Timur
aanpradana223@gmail.com

Aji Qolbu Wibawa Syahalam
Department of Information System, Faculty of Computer Science, UPN Veteran Jawa Timur
Surabaya, Jawa Timur
20082010176@student.upnjatim.ac.id

***Abstract*— The birth rate constitutes a significant metric within the field of demographics. It serves as a crucial element influencing the strategic planning and development of a region, particularly in provinces characterized by notably large populations, such as North Sumatra. The process of forecasting an optimal birth rate necessitates the involvement of multiple agencies and services to effectively devise policies pertaining to health, education, and enhanced infrastructure for the future. This study employs modeling techniques utilizing the Naive Bayes algorithm. This particular algorithm represents a probabilistic classification method within the realm of data mining, aimed at predicting birth rates across all districts and municipalities in North Sumatra, leveraging demographic and socio-economic datasets commencing from the year 2022. The dataset encompasses variables such as population statistics, demographics of women of reproductive age, levels of educational attainment, accessibility to health services, and incidences of poverty, all of which were sourced from the Central Statistics Agency (BPS) over a five-year timeframe. The research methodology is executed through several phases, including data preprocessing, feature selection, partitioning of training and test datasets, and a validation testing process to affirm the reliability of the proposed model. The dataset is partitioned into training and test components utilizing a distribution ratio of 70:30. The outcomes of the proposed model's testing are computed, employing a confusion matrix to derive metrics such as accuracy, precision, recall, and F1 scores. The results yield an accuracy value of 85%, a precision of 87%, and an F1 score of 86%. These findings indicate a favorable outcome in regional mapping and reflect an appropriate birth rate. Subsequently, the results are visualized within a geographic information system (GIS) to elucidate the spatial patterns of the predicted birth rate, thereby facilitating local government interventions in specific areas.**

Keywords*— *birth rate prediction, Naive Bayes, North Sumatra, demographic analysis, machine learning, visualization

I. INTRODUCTION

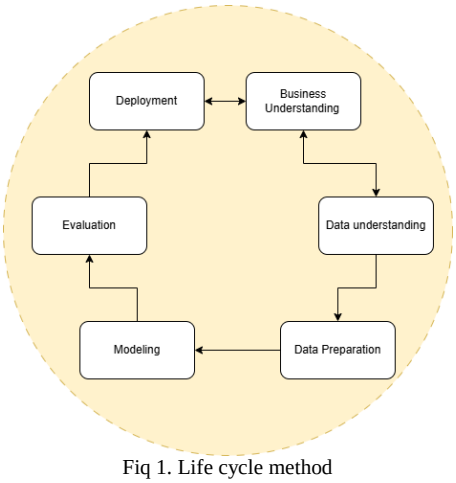
The population of a region is an important focus in development planning, including understanding and managing demographic aspects such as birth rates. The North Sumatra region as one of the provinces in Indonesia is used in this research to look further into birth rate trends, the results of which are used to design policies that are in accordance with the needs of its people. In the era of digital transformation, technology, and data analysis methods play an important role in understanding patterns and predicting social phenomena, including birth rates. The algorithm used is Naive Bayes, a classification method in machine learning, offering an effective approach to modeling the relationship between various factors and predicting birth rates. Birth rate prediction using the Naive Bayes algorithm in the North Sumatra region can look at previous research on similar applications in other regions. The Naive Bayes algorithm, a simple but effective classification technique, has been successfully used to predict birth rates in various contexts, such as in Salatiga City and Randau Jekak Village, by analyzing historical data and identifying patterns that can predict future trends [1] [2]. This approach can be adapted to North Sumatra, taking into account unique demographic and socio-economic factors. Application of Naive Bayes in Birth Rate Prediction. For example, the Naive Bayes algorithm is used to predict birth rates by analyzing data from 2019-2023, helping to manage population density issues [1]. The algorithm effectively classifies birth rates as high in 2021, helping local governments plan for population growth [2]. In previous studies related to North Sumatra research, North Sumatra has experienced rapid population growth influenced by fertility, mortality, and migration, which are crucial variables for prediction models and the success of Naive Bayes depends on the availability of comprehensive and accurate historical data, which can be challenging in areas with sparser census data [3] Previous studies Although Naive Bayes is effective, other algorithms such

as Neural Networks have shown higher accuracy in similar prediction tasks, such as preterm birth prediction [4]. This suggests that while Naive Bayes is useful, exploring other algorithms can improve prediction accuracy. Although the Naive Bayes algorithm offers a practical approach to predicting birth rates, its effectiveness in North Sumatra will depend on the quality of the data and the specific demographic factors of the region [12]. In addition, alternative algorithms such as Neural Networks can provide more accurate predictions, although they may require more complex data processing and computational resources [13].

This study aims to explore the potential use of the Naive Bayes algorithm in predicting birth rates in the North Sumatra region. By combining demographic data, health conditions, and socio-economic factors, we can develop a model that can help the government and other stakeholders anticipate future population needs. The importance of this study lies in its contribution to data-based development planning, which can improve the efficiency of resource allocation and align demographic policies with the real needs of the community. By leveraging the power of data analysis, we can shape responsive policy directions, providing positive impacts for the people of North Sumatra.

II. METHOD

The CRISP-DM (Cross-Industry Standard Process for Data Mining) framework is an established methodology extensively employed in data mining initiatives across various sectors. This framework encompasses six principal stages: [6] Business Understanding, which pertains to the identification of organizational objectives and requirements; [7] Data Understanding, which entails the preliminary collection and examination of data; [8] Data Preparation, which involves the processes of data cleansing, transformation, and attribute selection to construct the definitive dataset; [9] Modeling, which consists of the selection of algorithms and the training of models; [10] Evaluation, which refers to the analysis of model efficacy and alignment with business objectives; and [6] Deployment, which is the integration of the model within a functional operational context.



Fiq 1. Life cycle method

This methodology is characterized by its iterative and adaptable nature, permitting modifications at each phase to enhance the outcomes [14]. The CRISP-DM framework has gained prominence due to its provision of a structured approach from the inception of the project to the ultimate deployment, thereby facilitating efficient data-driven decision-making methodologies.

A. *Prediction*

Prediction is the process or act of predicting or estimating an event or value in the future based on available information and data. The main goal of prediction is to identify patterns or trends from historical data and use this information to make estimates or projections regarding future events. In the context of science, data analysis, and artificial intelligence, prediction often involves the use of mathematical or statistical models to process data.

B. *Dataset*

A dataset is a collection of data that is organized and arranged in such a way that it can be used for analysis, research, model development, or other purposes. Datasets can include various types of data, including text, images, sounds, numbers, and so on. Datasets are an important component in the world of data science, machine learning, and statistical analysis because the quality and quantity of good data greatly influence the results of the analysis or model being built. Method Data collection was carried out through a literature study. The dataset was obtained by collecting references and retrieving data via the official BPS North Sumatra website which has been publicly published at https://sumut.bps.go.id/

This research will look for the probability value of each hypothesis with a class (label) related to the birth rate. To determine the predicted birth rate in 2022, you can use population data, so data from BPS is needed to determine whether the birth rate is high or low, especially in the city of Medan. Classification methods are also needed to find a model that can distinguish classes from data with the aim of predicting the class of an object whose label is undefined. Therefore, using the Naive Bayes algorithm you can estimate future opportunities through previous experience so that the high or low birth rate in Medan City can be known [16].

TABLE I. POPULATION DATA OF NORTH SUMATRA PROVINCE 2022

No	Summary					
	City	***Population***	***Man***	***Women***	***Marriages***	***Births***
1	Nias	149249	72990	76259	15281	3589
2	Mandailing Natal	484874	241594	243280	52825	9469
3	Tapanuli Selatan	307312	154457	152855	37265	5730
4	Tapanuli Tengah	374734	188740	185994	39415	8528
5	Tapanuli Utara	318424	159102	159322	32916	6249

No	Summary					
	City	***Population***	***Man***	***Women***	***Marriages***	***Births***
6	Toba	212133	105874	106259	21496	3637
7	Labuhanbatu	508024	257802	250222	60602	10325
8	Asahan	787681	398595	389086	103208	13609
9	Simalungun	1021615	513380	508235	113701	14981
10	Dairi	315460	158077	157383	32134	5818
11	Karo	414429	205035	209394	48510	8035
12	Deli Serdang	1953986	983675	970311	269840	41886
13	Langkat	1039926	526020	513906	147299	18372
14	Nias Selatan	373674	187627	186047	32797	7518
15	Humbang Hasundutan	202299	101296	101003	20824	4278
16	Pakpak Barat	54609	27603	27006	6409	1189
17	Samosir	139337	69442	69895	13601	2468
18	Serdang Bedagai	667998	336597	331401	89658	11100
19	Batubara	416367	209540	206827	58441	8123
20	Padang Lawas Utara	267275	136317	130958	30877	7317
21	Padang Lawas	267275	134713	132562	36473	7162
22	Labuhanbatu Selatan	320324	163636	156688	41407	7876
23	Labuhanbatu Utara	390954	198522	192432	48957	7537
24	Nias Utara	150780	75004	75776	16134	3197
25	Nias Barat	91346	44485	46861	9342	1898
26	Sibolga	90366	45335	45031	9878	1635
27	Tanjungbalai	179748	91099	88649	22345	3506
28	Pematang Siantar	274056	135566	138490	27241	3589
29	Tebing Tinggi	177785	88549	89236	21109	9469
30	Medan	2494512	124231	1252199	224075	5730
31	Binjai	300009	150032	149977	35555	8528
32	Padang Sidempuan	231062	115038	116024	26228	6249
33	Gunungsitoli	137583	66938	70645	16237	3637

C. Classification

Naive Bayes is a classification algorithm based on Bayes' probability theorem. This algorithm is used to predict the class of data based on the features associated with that data. The name "Naive" (simple) comes from the simple assumption made by this algorithm regarding the independence between each pair of features, although in reality, the features may not be completely independent. The following is a complete explanation of Naive Bayes:

$$P(H|X) = \frac{P(X|H) \times P(H)}{P\,(X)} \quad (1)$$

Where:

X: Data with unknown class

H: Hypothesis data is a specific class P(H|X): Probability of hypothesis H based on conditions (posterior probability)

P(H): Probability of hypothesis H (prior probability)

P (X|H): Probability based on the conditions in the hypothesis

Naive Bayes classification can also be interpreted as a classification technique based on probability theory and Bayes' theorem, with the assumption that each variable or parameter that determines a decision is independent so that the existence of each variable has nothing to do with other attributes. The process of the Naive Bayes method is as follows

1) Calculate the probability value of each hypothesis with the existing classes (labels). "P(XK|Ci)"
2) Calculating the accumulated probability of each Class "P(X|Ci)"
3) Calculating the Value of P(X|Ci)×P(Ci) Determine the Class of the new Case

III. RESULT AND DISCUCCION

A. Calculate

Based on the dataset obtained, birth rate classification can be calculated using the Naive Bayes algorithm by inputting data on the total number of births, marriage rates, and birth rates. The following are the steps needed to predict the birth rate in Medan City

TABLE II. PREDICT CRITERIA

No	Criteria 1	Criteria 2	Criteria 3
1	under 2494512	under 224075	under 34633
2	above 2494512	above 224075	above 34633

Next step calculation value $P(\mathrm{X}K|Ci)$

1) P (criteria 1 = "under 2494512 "classification= "low")

   P criteria 1 = 32/33)

   P (criteria 1 = "above 2494512 "classification= "high")

   P (criteria 1 = 1/33)

2) P (criteria 1 = "under 224075") Klasifikasi = "low")

P (criteria 1 = 31/33)

P (criteria 1 = "above 224075 "classification = "high")

P (criteria 1 = 2/33)

3) P (criteria 1 = "under 34633 "classification = "low")

P (criteria 1 = 31/33)

P (criteria 1 = "above 34633 "classification = "high")

P (criteria 1 = 2/33)

And this process calculation value $P(C\text{i})$

1) P (Classification= "Birthrate") = high

Frequency of values above the average number from birth rate classification data.

$$Mean = \frac{\sum TOTAL\ DATA}{\sum FREKUENSI} = \frac{278100}{33} = 8427{,}27 \quad (1)$$

This results:

P(Classification="Birthrate") =9/33 falls into the high classification, while the result of P(Classification = "Birthrate") = 24/33 shows a low classification value.

B. Evaluation

The examination employing the confusion matrix in the context of the Prediction of Birth Rates Utilizing the Naive Bayes Algorithm within the North Sumatra Region serves to assess the efficacy of the classification model in categorizing birth rate classifications (for instance: high, medium, and low) [11]. The confusion matrix delineates the number of accurate predictions (True Positive and True Negative) alongside inaccurate predictions (False Positive and False Negative) predicated on the test dataset. From this matrix, evaluation metrics such as accuracy, precision, recall, and F1-score can be derived. These findings furnish an insight into the degree to which the Naive Bayes model is proficient in discerning data patterns and effectively classifying birth rates within the North Sumatra region [15]. The outcomes procured from the testing and evaluation process indicate an accuracy value of 85%, a precision of 87%, and an F1 measure of 86%.

IV. CONCLUSION

This research explores the potential of the Naive Bayes algorithm in the context of predicting birth rates in the North Sumatra Region. The research results show that this approach can be an effective instrument for understanding and forecasting demographic trends, providing valuable insights for future development planning.

By utilizing demographic data, health conditions and socioeconomic factors, the model produced from the Naive Bayes Algorithm is able to provide fairly accurate predictions regarding birth rates in this region. The advantage of this algorithm lies in its ability to handle a large number of variables and complex aspects, as well as its robustness to imperfect data.

This conclusion confirms that the implementation of the Naive Bayes Algorithm can be a valuable tool in supporting decision making regarding demographic policy. Accurate prediction results can help governments, health institutions, and other stakeholders to allocate resources more efficiently, ensure adequate health services, and design development programs that are responsive to population needs.

However, this research also recognizes that predictions are estimates and can be influenced by changes in environmental conditions and other factors that are difficult to fully predict. Therefore, it is necessary to continuously update the model to take into account changes in the social, economic, and demographic context.

Overall, this research makes an important contribution to the development of development planning strategies in North Sumatra and encourages further research to deepen our understanding of the factors that influence birth rates and how best to manage them for community welfare.

REFERENCES

[1] Saputri, A. D., & Hendry, H. (2024). Prediction of the Birth Rate of Babies at Regional Hospitals in Salatiga City Using the Naïve Bayes Algorithm. Advance Sustainable Science, Engineering and Technology (ASSET), 6(2).

[2] Hartono, M. K., & Hendry, H. (2022). Prediction of baby birth rate using naïve bayes classification algorithm in randau village. Jurnal Teknik Informatika, 3(4), 863–868.

[3] Wahyuni, R., & Simbolon, N. T. (2020). Penerapan pemodelan stokastik dengan metode polinom newton gregory maju dan polinom newton gregory mundur dalam memprediksi jumlah penduduk sumatera utara. 6(1), 70–77.

[4] Puspitasari, D., Ramanda, K., Supriyatna, A., Wahyudi, M., Sikumbang, E. D., & Sukmana, S. H. (2020). Comparison of Data Mining Algorithms Using Artificial Neural Networks (ANN) and Naive Bayes for Preterm Birth Prediction. 1–6.

[5] Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526-534.

[6] Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining (PAKDD) (pp. 29–39). Springer.

[7] Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., & Hernández-Orallo, J. (2020). *CRISP-DM twenty years later: From data mining processes to data science trajectories*. IEEE Transactions on Knowledge and Data Engineering, 34(5), 1–15.

[8] Ayele, W. Y. (2020). *Adapting CRISP-DM for idea mining: A data mining process for generating ideas using a textual dataset*. International Journal of Advanced Computer Science and Applications (IJACSA), 11(6), 20–28.

[9] Shimaoka, A. M., Soares, D. A., & Prado, A. F. do. (2024). *The evolution of CRISP-DM for data science: Methods, processes, and frameworks*. Journal of Data Science and Innovation, 9(1), 45–60.

[10] Pizoń, J., Michalak, K., & Niewczas, P. (2023). *Data engineering in CRISP-DM process production data – Case study*. Procedia Computer Science, 217, 1022–1029.

[11] Nurdin, A., Kartika, D. S. Y., & Najaf, A. R. E. (2024). Klasifikasi Penyakit Daun Tomat Dengan Metode Convolutional Neural Network Menggunakan Arsitektur Inception-V3. *JURNAL ILMIAH INFORMATIKA*, *12*(02), 114-119.

[12] RahmaPutri, M. R., Kartika, D. S. Y., & Wati, S. F. A. (2024). Klasifikasi Tingkat Kepuasan Pelanggan Sat & Sun: the Almeaty Service Menggunakan Naive Bayesklasifikasi Tingkat Kepuasan Pelanggan Sat & Sun: the Almeaty Service Menggunakan Naive Bayes. Jurnal Informatika dan Teknik Elektro Terapan, 12(3).

[13] Firsttama, R. A., Arifiyanti, A. A., & Kartika, D. S. Y. (2024). Analisis Sentimen Komentar Youtube Konferensi Tingkat Tinggi G20 Menggunakan Metode Naive Bayes. Jurnal Teknologi Dan Sistem Informasi Bisnis, 6(2), 282-285.

[14] Fernaldy, F. A., Arifiyanti, A. A., & Kartika, D. S. Y. (2025). KLASTERISASI TRACER STUDY ALUMNI UNIVERSITAS XYZ MENGGUNAKAN ALGORITMA K-MEANS. Jurnal Informatika dan Teknik Elektro Terapan, 13(1).

[15] Nurqoulby, F., Arifiyanti, A. A., & Kartika, D. S. Y. (2024). Analysis Sentiment Of Users Internet Service Providers In Indonesia On Social Media X Using Support Vector Machine. Data Science: Journal of Computing and Applied Informatics, 8(2), 88-95.

[16] Koeshardianto, M., Permana, K. E., Kartika, D. S. Y., & Setiawan, W. (2024, July). Classification of dry-beans using synthetic minority over-sampling technique and stochastic gradient boosting machines. In AIP Conference Proceedings (Vol. 3176, No. 1). AIP Publishing.