International Journal Of Health, Engineering And Technology (IJHET) *E-ISSN 2829 - 8683*
Volume 4, Number 3, September 2025, *Page. 188 - 194*
Email : editorijhess@gmail.com

Classification of Infertility Risk in Female Patients Based on Medical Record Data Using Naive Bayes Algorithm

Fahruzi Sirait [1)], Halimah Tusakdiyah Harahap[2)], Nadya Fitriani[3)], Rika Handayani[4)],Baginda Restu Al Ghazali[5)]
[1,5)] Information Systems, Faculty of Computer Science, Ika Bina Institute of Technology and Health
[2,3,4)]Midwifery, Faculty of Health Sciences, Ika Bina Institute of Technology and Health

*Corresponding Author
Email : fahruzi.sirait21@itkes-ikabina.ac.id

Abstract
Infertility is a reproductive health problem that has a significant impact globally, especially in developing countries such as Indonesia. This study aims to classify the risk of infertility in female patients at Rantauprapat Regional Hospital by utilizing the Naive Bayes algorithm based on electronic medical record data. The data used consisted of 500 medical records of female patients of childbearing age during the period 2019–2022, which had been processed and divided into training data (70%) and testing data (30%). The analysis and modeling process was carried out using the RapidMiner application without requiring programming skills. The results showed that the Naive Bayes model was able to classify the risk of infertility with an accuracy level of 86.7%, precision of 91.0%, recall of 93.2%, and F1-score of 92.1
Keywords: Infertility, Data mining, Naïve bayes, Electronic medical records, Rapidminer

INTRODUCTION

Infertility is one of the most significant reproductive health problems globally, especially in developing countries such as Indonesia. According to the World Health Organization (WHO), infertility is defined as the failure of a fertile couple to achieve pregnancy after having regular sexual intercourse for 12 months without using contraception. WHO data shows that around 8–12% of couples in the world experience infertility, and this figure tends to increase from year to year. In Indonesia, the prevalence of primary infertility in women is estimated to reach 10–15%, with the most cases found in the reproductive age group. This is a challenge for the health sector, considering that infertility can have significant psychological, social, and economic impacts on individuals and families.

Many factors can influence the risk of infertility in women, both internal and external factors. Internal factors include age, history of irregular menstruation, abnormalities in the reproductive organs such as endometriosis or polycystic ovary syndrome (PCOS), and hormonal disorders. Meanwhile, external factors can include a history of sexually transmitted infections, exposure to chemicals, unhealthy lifestyles, obesity, stress, and other environmental factors. Several studies have shown that a history of systemic diseases, such as diabetes mellitus or autoimmune diseases, can also increase the risk of infertility in women. Given the many factors that can influence, the process of identifying and classifying infertility risks is very important to support early detection efforts and appropriate treatment.

Currently, early detection of infertility is very necessary so that medical intervention can be carried out more quickly and precisely. However, the biggest challenge faced by hospitals and health facilities in Indonesia is the limited human resources and equipment, as well as the less-than-optimal utilization of comprehensive medical record data. On the other hand, advances in information technology in the health sector, especially in the utilization of electronic health records (EHR), have enabled the collection of patient medical record data in a more structured and systematic manner. The collected medical record data can be used as the main source of information to analyze and identify certain patterns related to infertility. However, the utilization of medical record data in health care facilities in Indonesia is currently still not optimal, especially in supporting data-driven clinical decision making.

One solution that can be applied is the use of data mining techniques to extract important information from very large amounts of medical record data. Data mining is a data analysis process with the aim of finding certain patterns or relationships hidden in large data sets (big data). In the context of reproductive health, data mining can be used to predict or classify the risk of infertility in female patients

based on various clinical attributes available in medical records. This method is very useful to help health workers in making clinical decisions quickly, accurately, and based on evidence.

One of the data mining algorithms widely used in the health sector is Naive Bayes. Naive Bayes is a probability-based classification algorithm that applies the principle of Bayes' Theorem with the simple assumption that each attribute is independent of each other. The main advantage of Naive Bayes is its ability to provide fast, easy-to-interpret, and fairly accurate prediction results, even on complex health data. Various studies have shown the effectiveness of using the Naive Bayes algorithm in disease risk classification, for example in detecting heart disease, breast cancer, and diabetes. However, until now there has been very limited research that specifically utilizes this algorithm for infertility risk classification based on patient medical record data in regional hospitals, especially at Rantauprapat Regional Hospital. In fact, optimizing the use of medical record data in hospitals is very important to improve the quality of reproductive health services, especially in efforts to detect infertility in women early.

To support optimal data analysis without requiring programming skills, various data science platforms are now available such as RapidMiner. RapidMiner is one of the popular software used to perform data analysis, data preprocessing, and classification modeling and model evaluation without the need to write program code. This platform supports visual data analysis processes through drag and drop features, making it very suitable for use by health workers or researchers who do not have a programming background. Thus, the integration of the Naive Bayes algorithm in RapidMiner for infertility risk classification at Rantauprapat Regional Hospital is expected to provide an effective and efficient solution in supporting data-based clinical decision making.

Based on the description above, this study aims to: (1) apply the Naive Bayes algorithm to medical record data of female patients at Rantauprapat Regional Hospital for infertility risk classification, (2) evaluate the accuracy of the resulting classification model, and (3) identify the main factors that contribute to the risk of infertility in women. With this research, it is hoped that in the future it can provide a real contribution in the utilization of information technology in the field of obstetrics, especially in efforts to detect infertility risk early effectively and efficiently, and become an important scientific reference for the development of similar research in the future, both at the local and national levels.

RESEARCH METHODS

Types of research

This research is an applied research with a quantitative approach and case study design. The research focuses on the application of data mining methods to classify the risk of infertility in female patients based on medical record data at Rantauprapat Regional Hospital.

Data Collection Sources and Techniques

In this study, the amount of data used was 500 medical records of female patients of childbearing age (15–49 years) who underwent infertility-related examinations at Rantauprapat Regional Hospital during the period 2019–2022. The data used has gone through a selection process based on inclusion criteria, namely patient data with complete information on all research variables. Furthermore, the data was divided into 70% training data (350 data) and 30% testing data (150 data) for the modeling and evaluation process using the Naive Bayes algorithm in RapidMiner.

Table 1. Research Data Attributes

No.	Nama Variabel	Jenis Data	Deskripsi/Penjelasan

1	Usia Pasien	Numerik	Umur pasien dalam tahun
2	Lama Menikah	Numerik	Durasi pernikahan (tahun)
3	Riwayat Menstruasi	Kategorik	Teratur/Tidak Teratur
4	Riwayat Penyakit Reproduksi	Kategorik	Misal: PCOS, endometriosis, dll
5	Riwayat Infeksi Menular Seksual	Kategorik	Ya/Tidak
6	Riwayat Penyakit Sistemik	Kategorik	Misal: Diabetes, autoimun, dll
7	Hasil Pemeriksaan Hormonal	Numerik	Nilai FSH, LH, prolaktin, dsb (bisa lebih dari 1)
8	Indeks Massa Tubuh (IMT)	Numerik	Hasil perhitungan IMT
9	Gaya Hidup	Kategorik	Merokok, konsumsi alkohol, olahraga, dll
10	Status Infertilitas	Kategorik	Infertil/Tidak Infertil/Rendah/Sedang/Tinggi

Research Stages

The research was conducted through several stages in Table 2 as follows:

TABLE 2. Research Stages

No.	Tahapan Penelitian	Penjelasan Singkat
1	Pengumpulan Data Rekam Medis (2019–2022)	Mengambil data pasien dari RSUD Rantauprapat sesuai kriteria inklusi penelitian
2	Pembersihan & Pra-pemrosesan Data	Seleksi, pembersihan, penanganan data hilang, dan transformasi data
3	Pembagian Data Training dan Testing	Split data untuk training (70%) dan testing (30%) menggunakan RapidMiner
4	Penerapan Algoritma Naive Bayes	Membangun model klasifikasi menggunakan RapidMiner
5	Evaluasi Model	Menghitung akurasi, precision, recall, F1-score, dan confusion matrix
6	Analisis Faktor Risiko Utama	Mengidentifikasi atribut paling berpengaruh menggunakan feature importance
7	Interpretasi Hasil & Penarikan Kesimpulan	Menyimpulkan temuan penelitian untuk rekomendasi bidang kebidanan

Research Flowchart

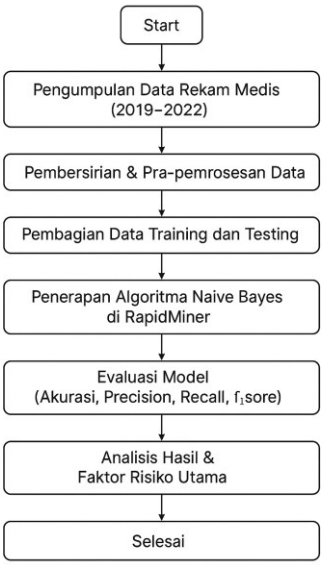

FIGURE 1. Research Flowchart

The flowchart illustrates the research stages in classifying infertility risk using the Naive Bayes algorithm. The research began with the collection of medical record data of female patients for the period 2019–2022, followed by the data cleaning and pre-processing process to ensure the data was ready for analysis. After that, the data was divided into training and testing data. The Naive Bayes algorithm was then applied using RapidMiner to build a classification model. The resulting model was evaluated by measuring accuracy, precision, recall, and f1-score. Finally, an analysis of the model results and identification of the main risk factors was carried out before the research was declared complete.

Research Ethics

This study has obtained official permission from Rantauprapat Regional Hospital. All patient data is processed anonymously, without including the patient's personal identity, and complies with the principles of research ethics and patient data protection.

RESULTS AND DISCUSSION

Naive Bayes Classification Results

The infertility risk classification process is carried out by dividing the data into 70% training data (350 data) and 30% testing data (150 data). The Naive Bayes model is implemented using the RapidMiner application. The results of the model performance evaluation on the testing data are shown in Table 1 below:

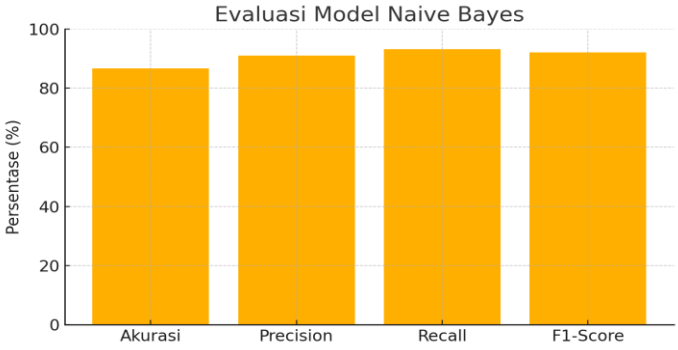

Figure 2. Naive Bayes Model Evaluation Results

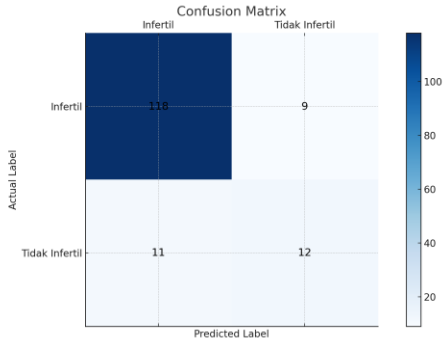

Figure 3. Confusion Matrix Prediction Results

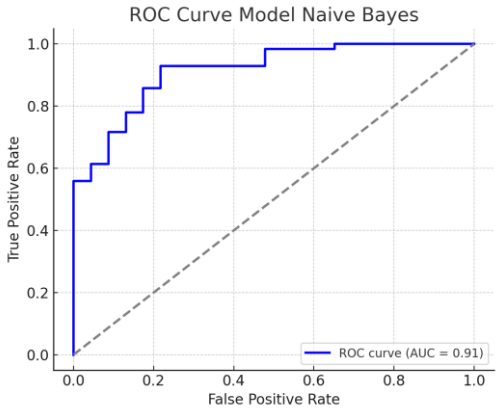

Figure 4. ROC Curve

Key Risk Factor Analysis

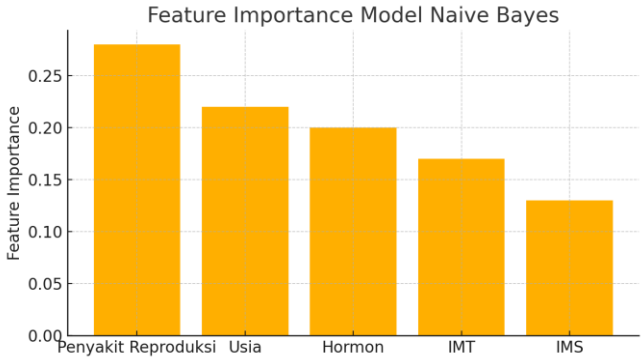

Figure 5. Analysis of Key Risk Factors

Based on the feature importance analysis on the Naive Bayes algorithm, it was found that several main factors that most influence the classification of infertility risk in female patients at Rantauprapat Hospital are:

1) History of Reproductive Diseases (such as PCOS, endometriosis)
2) Patient age
3) Hormonal Examination Results (FSH, LH, Prolactin)
4) Body Mass Index (BMI)
5) History of Sexually Transmitted Infections (STIs)

These factors play a significant role in determining the risk of infertility, in accordance with findings in previous studies. Patients with a history of reproductive diseases, age over 35 years, hormonal imbalance, abnormal BMI, and history of STIs tend to have a higher risk of infertility.

Model Evaluation

Model evaluation was conducted to assess the performance of the Naive Bayes algorithm in classifying the risk of infertility in female patients. The evaluation was conducted by comparing the results of the model predictions to the testing data (as many as 150 data or 30% of the total data). Several evaluation metrics used include accuracy, precision, recall, and F1-score.

The evaluation results showed that the model had an accuracy rate of 86.7%, which means that the model can predict infertility status correctly on most of the testing data. A precision value of 91.0% indicates that most of the patients predicted as "Infertile" by the model are indeed infertile. Meanwhile, a recall value of 93.2% indicates that the model successfully identified almost all patients who were truly infertile. The F1-score value of 92.1% represents the balance between precision and recall achieved by the model.

However, the confusion matrix results show that there is still a small amount of misclassified data, either in the form of false positives or false negatives. This can be caused by similarities in characteristics between infertile and non-infertile patients or other variables that have not been recorded in the data.

In general, the performance of the Naive Bayes model is good enough to be used in early detection of infertility risk based on medical record data, although there is an opportunity to improve accuracy by increasing the amount of data, clinical attributes, or applying data balancing techniques.

CONCLUSION

Based on the results of the research that has been conducted, it can be concluded that: The Naive Bayes algorithm applied to medical record data of female patients at Rantauprapat Regional Hospital for the period 2019–2022 is able to classify infertility risk with an accuracy rate of 86.7%. The factors that most influence the classification of infertility are history of reproductive disease, patient age, hormonal examination results, body mass index (BMI), and history of sexually transmitted infections.

REFERENCES

A. F. Butts and C. D. Shaughnessy, “Environmental and lifestyle risk factors for infertility,” Best Practice & Research Clinical Obstetrics & Gynaecology, vol. 35, pp. 3–10, 2016.

A. Jayanthi and S. Subashini, “Classification of Diabetes Disease Using Support Vector Machine,” International Journal of Computer Applications, vol. 62, no. 12, pp. 11–14, 2013.

A. Otoom and E. E. Abdallah, “Breast cancer diagnosis system based on Naive Bayes classifier,” International Journal of Computer Applications, vol. 99, no. 10, pp. 13–17, 2014.

B. Aldosari, Health Informatics: Practical Guide for Healthcare and Information Technology Professionals, 7th ed., Lulu.com, 2017.

C. J. Cousineau and A. Domar, “Psychological impact of infertility,” Best Practice & Research Clinical Obstetrics & Gynaecology, vol. 21, no. 2, pp. 293–308, 2007.

D. Novitasari, S. Setyawati, and F. Ramadani, “Pemanfaatan Rekam Medis Elektronik untuk Kesehatan Reproduksi,” Jurnal Kesehatan Masyarakat, vol. 14, no. 2, pp. 121–128, 2019.

G. M. Joffe, "Lifestyle and environmental contributions to male infertility," BMJ, vol. 336, no. 7644, pp. 1403–1404, 2008.

H. Kaur and S. K. Wasan, "Empirical Study on Applications of Data Mining Techniques in Healthcare," Journal of Computer Science, vol. 2, no. 2, pp. 194–200, 2006.

H. Zhang and C. X. Ling, "An improved learning algorithm for naive Bayes," in Proc. 17th Int. Joint Conf. Artif. Intell., 2001, pp. 903–908.

J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2012.

M. A. Fritz and L. Speroff, Clinical Gynecologic Endocrinology and Infertility, 8th ed., Philadelphia: Lippincott Williams & Wilkins, 2011.

M. Hofmann and R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, CRC Press, 2016.

M. N. Mascarenhas, S. R. Flaxman, T. Boerma, S. Vanderpoel, and G. Stevens, "National, regional, and global trends in infertility prevalence since 1990: a systematic analysis of 277 health surveys," PLoS Med., vol. 9, no. 12, p. e1001356, 2012.

M. T. O. Widyaningsih and E. N. Paramita, "Analisis Data Rekam Medis Elektronik untuk Prediksi Risiko Kehamilan Menggunakan Data Mining," Jurnal Kesehatan Reproduksi, vol. 11, no. 2, pp. 123–130, 2020.

N. S. Checa Vizcaíno, F. Vassena, and S. Rodríguez, "Infertility: impact of lifestyle and environmental factors," Gynecological Endocrinology, vol. 34, no. 7, pp. 473–477, 2018.

Perkumpulan Obstetri dan Ginekologi Indonesia (POGI), "Pedoman Nasional Pelayanan Kedokteran Tata Laksana Infertilitas," 2022.

Practice Committee of the American Society for Reproductive Medicine, "Diagnostic evaluation of the infertile female: a committee opinion," Fertil. Steril., vol. 103, no. 6, pp. e44–e50, 2015.

R. Kurniawan, A. Subekti, and E. S. Putra, "Analisis Efektivitas RapidMiner dalam Klasifikasi Data Kesehatan," Jurnal Informatika dan Komputer, vol. 6, no. 2, pp. 99–108, 2021.

S. J. Delaney, J. S. Johnston, and G. D. Leith, "The application of naive Bayes classification to the prediction of heart disease," Artificial Intelligence in Medicine, vol. 27, no. 1, pp. 53–61, 2003.

S. W. Nugroho, R. Astuti, and N. K. Yuniar, "Prevalensi Infertilitas di Indonesia," Jurnal Kesehatan Reproduksi, vol. 8, no. 2, pp. 95–102, 2017.

Suyanto, "Data Mining untuk Klasifikasi Data Medis," Jurnal Informatika, vol. 12, no. 1, pp. 23–34, 2018.

V. Kotu and B. Deshpande, Data Science: Concepts and Practice, Morgan Kaufmann, 2019.

W. Ombelet, J. Cooke, J. Dyer, Z. Serour, and G. Devroey, "Infertility and the provision of infertility medical services in developing countries," Hum. Reprod. Update, vol. 14, no. 6, pp. 605–621, 2008.

World Health Organization, "Infertility," 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/infertility [Accessed: 29-Apr-2024].

Y. Jiang and Y. Yao, "Naive Bayes Classification," in Encyclopedia of Machine Learning and Data Mining, Springer, 2017, pp. 892–896.