Analisis Klasifikasi Kanker Payudara Menggunakan Algoritma Naive Bayes

Hardian Oktavianto[1], Rahman Puji Handri[2]
[1,2] Program Studi Teknik Informatika, Universitas Muhammadiyah Jember
[1]hardian@unmuhjember.ac.id, [2]pujihandri@gmail.com

ABSTRACT

Breast cancer is one of the highest causes of death among women, this disease ranks second cause of death after lung cancer. According to the world health organization, 1 million women get a diagnosis of breast cancer every year and half of them die, in general this is due to early treatment and slow treatment resulting in new cancers being detected after entering the final stage. In the field of health and medicine, machine learning-based classification has been carried out to help doctors and health professionals in classifying the types of cancer, to determine which treatment measures should be performed. In this study breast cancer classification will be carried out using the Naive Bayes algorithm to group the types of cancer. The dataset used is from the Wisconsin breast cancer database. The results of this study are the ability of the Naive Bayes algorithm for the classification of breast cancer produces a good value, where the average percentage of correctly classified data reaches 96.9% and the average percentage of data is classified as incorrect only 3.1%. While the level of effectiveness of classification with naive bayes is high, where the average value of precision and recall is around 0.96. The highest precision and recall values are when the test data uses a percentage split of 40% with the respective values reaching 0.974 and 0.973.

Keyword: Classification, Breast cancer, Naive bayes

1. Introduction

Kanker payudara merupakan salah satu faktor penyebab kematian tertinggi di kalangan wanita, penyakit ini menempati urutan kedua penyebab kematian setelah kanker paru – paru [7]. Faktor – faktor yang dapat memicu tumbuhnya kanker payudara seperti penuaan, jenis kelamin, ras, riwayat keluarga, genetika, atau perilaku pribadi seperti merokok, minuman beralkohol, dan diet [3]. Menurut organisasi kesehatan dunia, 1 juta wanita didiagnosis menderita kanker payudara setiap tahun dan separuh dari mereka akhirnya meninggal, pada umumnya hal ini disebabkan penanganan dini serta pengobatan yang lambat mengakibatkan kanker baru terdeteksi setelah memasuki stadium akhir [1].

Machine learning merupakan bagian dari bidang kecerdasan buatan yang mempunyai fokus pada penerapan algoritma maupun metode tertentu untuk melakukan prediksi, pengenalan pola, dan klasifikasi [2]. Pada umumnya tahapan pada *machine learning* adalah koleksi data, pembentukan model, tahap pelatihan, serta tahap uji. Pada bidang kesehatan dan pengobatan, klasifikasi yang berbasis *machine learning* telah banyak dilakukan untuk membantu dokter dan ahli kesehatan dalam mengelompokkan jenis tumor, hingga menentukan tindakan pengobatan yang sebaiknya dilakukan [1].

Pada penelitian ini akan dilakukan klasifikasi kanker payudara dengan menggunakan algoritma Naive Bayes untuk melakukan pengelompokan jenis kanker, apakah jinak atau ganas. Dataset yang digunakan adalah berasal dari basis data kanker payudara Wisconsin. Basis data ini terdiri dari 699 sampel yang diambil dari *Fine Needle Aspirates* (FNA) jaringan payudara manusia. Hasil dari penelitian ini diharapkan dapat menunjang dunia kesehatan khususnya pada kasus kanker payudara dengan menyediakan model klasifikasi yang dapat digunakan untuk melakukan prediksi maupun pengambilan keputusan medis lainnya.

2. Research Method

Tahapan penelitian kali ini secara garis besar ada 4 tahapan, yaitu : Studi Literatur, Praproses Data, Klasifikasi, dan Analisis.

Gambar 1. Tahapan Penelitian

Studi literatur dilakukan pertama kali sebagai tahapan persiapan terhadap keseluruhan penelitian, tahapan ini bertujuan untuk mencari dasar – dasar teori yang digunakan serta mencari pustaka – pustaka ilmiah yang mendukung, juga melakukan persiapan data. Tahapan selanjutnya adalah praproses data, yaitu melakukan persiapan terhadap set data yang telah diperoleh agar siap untuk dilakukan proses klasifikasi. Tahapan klasifikasi dilakukan dengan menggunakan alat bantu WEKA, sebuah perangkat lunak *machine learning* dan *data mining* yang sering digunakan pada penelitian – penelitian untuk melakukan uji algoritma maupun metode. Tahapan terakhir yaitu melakukan analisis atau penarikan kesimpulan terhadap hasil uji klasifikasi yang telah dilakukan pada tahap sebelumnya.

2.1 Jenis Data dan Sumber Data

Dataset yang digunakan adalah data kanker payudara Wisconsin yang diakses dari https://archive.ics.uci.edu/ml/. Dataset ini mempunyai 699 buah *record* dengan 11 buah variabel, yaitu : *id number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses,* dan *Class*.

Sebelum data digunakan untuk tahapan klasifikasi, dataset akan dipersiapkan terlebih dahulu. Adapun beberapa proses yang dilakukan nantinya meliputi : pembersihan data yang tidak lengkap (*missing value*), transformasi data, dan konversi data.

2.2 Klasifikasi

Klasifikasi pada penelitian ini menggunakan WEKA, sebuah perangkat lunak *open source* yang telah dikenal dan umum digunakan di bidang *data mining* dan *machine learning*. Pada WEKA banyak tersedia *library – library* algoritma, baik untuk klasifikasi, *clustering*, *association rule*, bahkan sampai *library* untuk praproses data dan *feature selection*.

Klasifikasi nantinya akan menerapkan algoritma Naive Bayes, dengan skenario uji dataset menggunakan pembagian persentase (*percentage split*), yaitu membagi jumlah keseluruhan data menjadi data latih dan data uji, sebagai contoh, apabila digunakan *percentage split* 60% berarti 60% dari total jumlah data akan digunakan sebagai data latih sedangkan sisanya sebanyak 40% akan digunakan sebagai data uji. Adapun *percentage split* yang dipakai adalah 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, dan 90%, hal ini dilakukan untuk mendapatkan hasil yang obyektif dari *percentage split* yang berbeda – beda.

Teorema Bayes dirumuskan sebagai : $p(h|d) = \frac{p(d|h)p(h)}{p(d)}$

dimana *h* merupakan label kelas yang menjadi target klasifikasi dan *d* merupakan variabelnya. Teorema Bayes dapat diartikan sebagai pembelajaran berdasarkan data latih untuk pembangunan model dari setiap kombinasi *h* dengan semua fitur *d* sehingga kita akan mendapatkan informasi peluang perolehan kelas *h* berdasarkan variabel – variabel *d* yang diamati. Setelah model dibangun, maka model tersebut digunakan sebagai dasar dari penentuan label untuk kelas data uji. Penentuan label kelas ini berdasarkan nilai peluang terbesar terhadap masing – masing kelas target.

2.3 Analisis

Tahap analisis adalah tahap terakhir dari keseluruhan penelitian, dimana akan dilakukan analisis terhadap hasil dari uji pada tahap klasifikasi yang telah dilakukan sebelumnya. Analisis dilakukan agar mengetahui sejauh mana hasil klasifikasi berdasarkan skenario uji yang telah dirancang. Beberapa nilai yang dibandingkan adalah : Jumlah data yang diklasifikasi dengan benar, jumlah data yang diklasifikasi salah, *Precision*, dan *Recall*.

3. Result and Analysis

Pada bab ini akan disampaikan mengenai hasil dari penelitian yang telah dilakukan. Pada sub bab pertama akan menjelaskan tentang praproses data, dimana dilakukan pembersihan data, transformasi data, dan konversi data. Kemudian dilanjutkan dengan proses klasifikasi dengan menggunakan *percentage split* yang berbeda, yaitu 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, dan 90%. Proses selanjutnya adalah melakukan analisis terhadap masing – masing hasil klasifikasi yang mencakup berapakah jumlah data yang diklasifikasi dengan benar, jumlah data yang diklasifikasi salah, nilai *Precision*, dan nilai *Recall*.

3.1. Praproses Data

Pada dataset ditemukan 16 buah data tidak mempunyai nilai variabel yang lengkap sehingga data – data tersebut dihapus. Dari total 699 data dan setelah dilakukan penghapusan 16 buah data maka total jumlah data menjadi 683 buah. Selain dilakukan penghapusan 16 buah data, juga dilakukan variabel yang memang

tidak dibutuhkan pada saat proses klasifikasi, yaitu variabel *id number*, sehingga pada dataset terdapat 10 variabel dari yang sebelumnya 11 variabel. Setelah pembersihan data, maka langkah selanjutnya adalah transformasi dan konversi data, hal ini dilakukan agar dataset dapat dibaca dan kemudian diolah melalui perangkat lunak WEKA. Langkah transformasi data yang dilakukan adalah merubah format atau susunan data tanpa merubah format file, jadi setelah dilakukan transformasi data maka format file tetap berupa format teks. Konversi data merupakan langkah terakhir pada praproses data, yaitu merubah format atau melakukan konversi format file dari teks ke bentuk *comma separated value* (.csv) sehingga nantinya dapat dibaca oleh perangkat lunak WEKA. Format file csv ini mirip dengan format file *spreadsheet* pada umumnya, yaitu terdapat baris dan kolom.

3.2. Klasifikasi

Total uji klasifikasi yang dilakukan pada penelitian ini adalah 10 kali, dengan variasi *percentage split* yang berbeda – beda. Pada tabel 1 menyajikan hasil uji klasifikasi dengan rincian persentase data terklasifikasi benar dan persentase data terklasifikasi salah. Nilai rata – rata persentase data terklasifikasi benar adalah 96.9% dengan nilai persentase tertinggi adalah 97.3% dan persentase terendah adalah 96%, sedangkan untuk rata – rata persentase data terklasifikasi salah adalah 3.11% dengan persentase tertinggi adalah 4.01% dan persentase terandah adalah 2.68%. Berdasarkan tabel 2 maka klasifikasi dengan algoritma naive bayes ini, secara rata – rata menghasilkan performa yang baik dimana rata – rata persentase data yang terklasifikasi dengan benar mencapai 96.9% dan rata – rata persentase data terklasifikasi salah hanya 3.11%.

Tabel 1 Hasil Uji Coba

Uji ke-	Percentage Split	Jumlah Data Latih	Jumlah Data Uji	Persentase Data Terklasifikasi Benar	Persentase Data Terklasifikasi Salah
1	5%	34	649	96.0%	4.01%
2	10%	68	615	97.2%	2.76%
3	20%	137	546	97.1%	2.93%
4	30%	205	478	97.1%	2.93%
5	40%	273	410	97.3%	2.68%
6	50%	342	341	97.1%	2.93%
7	60%	410	273	96.7%	3.30%
8	70%	478	205	97.1%	2.93%
9	80%	546	137	96.4%	3.65%
10	90%	615	68	97.1%	2.94%

Selain memperhatikan hasil uji dari persentase data terklasifikasi benar atau salah, maka pada penelitian ini juga digunakan *precision* dan *recall* untuk mengukur kinerja penerapan naive bayes terhadap klasifikasi kanker payudara. Precision digunakan untuk mengukur tingkat keberhasilan dari kelas data positif yang diklasifikasi dengan benar dari keseluruhan hasil klasifikasi kelas positif. *Recall* digunakan untuk menunjukkan tingkat keberhasilan dari kelas data positif yang diklasifikasikan benar dari keseluruhan data kelas positif. Atau bisa juga dikatakan bahwa *precision* mengukur kualitas klasifikasi sedangkan *recall* mengukur kuantitas klasifikasi. Pada penelitian ini data positif adalah data kelas kanker payudara jinak, sedangkan data negatif adalah data kelas kanker payudara ganas.

Tabel 2 Nilai TP Rate, FP Rate, Precision, dan Recall

Uji ke-	Percentage Split	Jumlah Data Latih	Jumlah Data Uji	TP Rate	Precision	Recall

1	5%	34	649	0.96	0.96	0.96
2	10%	68	615	0.972	0.972	0.972
3	20%	137	546	0.971	0.971	0.971
4	30%	205	478	0.971	0.971	0.971
5	40%	273	410	0.973	0.974	0.973
6	50%	342	341	0.971	0.972	0.971
7	60%	410	273	0.967	0.968	0.967
8	70%	478	205	0.971	0.971	0.971
9	80%	546	137	0.964	0.964	0.964
10	90%	615	68	0.971	0.971	0.971

Tabel 2 menyajikan rekapitulasi nilai *TP Rate, FP Rate, Precision*, dan *Recall* untuk masing – masing uji berdasarkan *percentage split*. Nilai rata – rata untuk *TP Rate, FP Rate, Precision,* dan *Recall* secara berurutan adalah 0.96, 0.03%, 0.96, dan 0.96. Berdasarkan hasil tersebut maka dapat kita ketahui bahwa rata – rata *precision* dan *recall* berada di sekitar 0.96 dimana mendekati nilai maksimal yaitu 1 sehingga tingkat efektivitas klasifikasi dengan naive bayes ini termasuk tinggi. Nilai *precision* dan *recall* paling tinggi yaitu pada uji dengan menggunakan *percentage split* 40% dengan nilai masing – masing 0.974 dan 0.973.

4. Conclusion

Kesimpulan dari penelitian ini adalah performa algoritma naive bayes untuk klasifikasi kanker payudara menghasilkan nilai yang baik, dimana rata – rata persentase data yang terklasifikasi dengan benar mencapai 96.9% dan rata – rata persentase data terklasifikasi salah hanya 3.1%. Sedangkan tingkat efektivitas klasifikasi dengan naive bayes ini termasuk tinggi, dimana rata – rata nilai *precision* dan *recall* berada di sekitar 0.96. Nilai *precision* dan *recall* paling tinggi yaitu ketika data uji menggunakan *percentage split* 40% dengan nilai masing – masing secara berurutan mencapai 0.974 dan 0.973.

Saran terhadap penelitian ini adalah penerapan naive bayes untuk klasifikasi ini masih bisa dikembangkan untuk lebih menguji performanya, yaitu dengan cara menerapkan pada dataset yang bersifat tidak seimbang atau *imbalanced*, selain itu pengembangan juga masih dapat dilakukan dengan menggabungkan naive bayes dengan metode seleksi fitur untuk lebih meningkatkan performa hasil klasifikasi.

References

[1] Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast Cancer Classification Using Machine Learning. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2018. https://doi.org/10.1109/EBBT.2018.8391453.

[2] Bazazeh, D., & Shubair, R. (2017). Comparative Study Of Machine Learning Algorithms For Breast Cancer Detection And Diagnosis. In International Conference on Electronic Devices, Systems, and Applications. https://doi.org/10.1109/ICEDSA.2016.7818560.

[3] Eleyan, A. (2012). Breast Cancer Classification Using Moments, 235(42003).

[4] Han J, Kamber M. 2001. Data Mining Concepts & Techniques. USA (US): Academic Press.

[5] Larose DT. 2005. Discovering Knowledge in Data : An Introduction to Data Mining. Canada (US) : John Wiley & Sons, Inc.

[6] Mitsa, T. 2010. Data Mining and Knowledge Discovery Series. Minneapolis (US): Chapman & Hall/CRC.

[7] Shah, C., & Jivani, A. G. (2013). Comparison Of Data Mining Classification Algorithms For Breast Cancer Prediction. In 2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013. https://doi.org/10.1109/ICCCNT.2013.6726477.

[8] Tan PN, Steinbach M, Kumar V. 2006. Introduction to Data Mining. Boston (US): Pearson Education.

[9] Witten IH, Frank E, Hall MA. 2011. Practical Machine Learning Tools and Techniques. San Fransisco (US) : Morgan Kauffman.