Analisis Kualitas Data dan Klasifikasi Data Pasien Kanker

Ahmad Fathan Hidayatullah[1], Alan Dwi Prasetyo[2], Dantik Puspita Sari[3], Intan Pratiwi[4]

Jurusan Teknik Informatika Universitas Islam Indonesia
Jl. Kaliurang km 14 Yogyakarta 55510
Telp (0274) 895287 ext 122, fax (0274) 895007 ext 148
fathan@uii.ac.id[1], alandwiprasetyo@gmail.com[2]
naylah.azifah.syahirah@gmail.com[3], tanintanpratiwi@gmail.com[4]

Abstract. Masalah yang ditemui dalam *dataset* yang besar adalah adanya duplikasi data dan *missing value*. Duplikasi terjadi karena ada perbedaan identifikasi antara entitas yang sama dalam dunia nyata misalnya duplikasi data pasien rumah sakit. Solusi dari permasalahan duplikasi adalah dengan melakukan deduplikasi. Deduplikasi dilakukan dengan mengeliminasi data yang memiliki kemiripan. Pendeteksian duplikasi data dilakukan dengan *Algoritma Levenshtein distance. Missing value* terjadi jika ada nilai dari suatu atribut yang tidak ditemukan. Atribut yang mengandung *missing value* diganti dengan nilai rata-rata seluruh data dalam setiap atribut. Setelah duplikasi data dan *missing value* dapat diatasi, kemudian dilakukan klasifikasi untuk mengidentifikasi adanya kesamaan data. Klasifikasi dilakukan dengan *tools* WEKA menggunakan algoritma *Decision Tree* dan *Naive Bayes*. Metode *Decision tree* menghasilkan akurasi sebesar 99.9988 % sedangkan metode *Naive Bayes* menghasilkan akurasi 99.9799 %. Akurasi yang diperoleh algoritma *Decision Tree* memiliki hasil sedikit lebih baik daripada *Naive Bayes*. Namun demikian, secara umum metode *Decision Tree* dan *Naive Bayes* sama-sama memiliki akurasi yang baik dalam melakukan klasifikasi kemiripan data pasien.

Keywords: duplikasi data, *missing value*, klasifikasi, *Decision Tree, Naive Bayes*

1 Pendahuluan

Data pasien merupakan data yang sangat penting dalam dunia kesehatan. Data pasien yang disimpan secara terstruktur dapat memberikan informasi tentang riwayat penyakit pasien. Namun demikian, ada beberapa kendala yang dihadapi untuk memperoleh informasi tersebut terkait dengan *dataset* pasien yang cukup besar. Diantara permasalahan yang ditemui dalam *dataset* pasien tersebut adalah adanya duplikasi data dan adanya *missing value*.

Duplikasi data dalam *data mining* dapat terjadi karena dua hal yaitu adanya *record* yang berulang dan adanya perbedaan identifikasi antara entitas yang sama dalam dunia nyata[1]. Adanya duplikasi data pada *dataset* dapat mempengaruhi kualitas performa *data mining*[2]. Duplikasi data pada *dataset* pasien dimungkinkan terjadi karena *input* data pasien dilakukan oleh orang atau waktu yang berbeda. Selain itu, duplikasi juga bisa terjadi karena adanya kesalahan penulisan pada saat proses input data se-

hingga terdapat beberapa data yang memiliki kemiripan atau bahkan sama dalam *dataset*.

Missing value dalam *dataset* pasien berasal dari data-data yang atributnya tidak memiliki nilai. Informasi ini tidak diperoleh dimungkinkan karena adanya data pasien yang tidak lengkap seperti jenis kelamin, nama belakang pasien, tanggal lahir pasien, dan sebagainya. Keberadaan *missing value* ini juga dapat menyebabkan duplikasi data karena ada lebih dari satu data dengan nama yang sama dan memiliki kelengkapan data yang berbeda.

Dalam hal ini, diperlukan adanya *record linkage* yaitu terhubungnya data satu pasien dengan pasien yang lain. *Record linkage* diperlukan apabila pihak medis memerlukan informasi tentang riwayat penyakit dari pasien sedangkan pasien pernah ditangani di beberapa tempat berbeda. Data riwayat penyakit dimungkinkan dapat diperoleh apabila ada hubungan antar data dari beberapas sumber[3]. Hal ini akan memberikan kemudahan bagi pihak medis, pasien, dan peneliti untuk mengetahui riwayat penyakit pasien[4].

Berdasarkan beberapa masalah yang telah dipaparkan, penelitian ini mencoba menyelesaikan masalah duplikasi data dan *missing value* pada data pasien kanker. Data tersebut diperoleh dari tahun 2005 sampai 2008. Selain itu, dalam penelitian ini dilakukan klasifikasi data pasien yang diprediksi memiliki kemiripan satu sama lain. Klasifikasi dilakukan dengan metode *Decision Tree* dan *Naive Bayes*.

2 Landasan Teori

2.1 Pendeteksian Duplikasi Data

Pendeteksian duplikasi data merupakan proses identifikasi *record* yang berbeda yang mengacu pada satu entitas atau objek yang memiliki kesamaan dalam dunia nyata[5]. Penelitian ini mendeteksi duplikasi data menggunakan *Algoritma Levenshtein distance*. Algoritma *Levenshtein distance* merupakan salah satu algoritma *Approximate String Matching* yang digunakan dalam pencarian string berdasarkan pendekatan perkiraan[6]. *Levenshtein distance* dilakukan untuk mencari kesamaan antara dua buah *string*[7]. Rumus *Levenshtein distance* direpresentasikan oleh persamaan (1).

$$s(v,w) = 1 - \frac{d(v,w)}{max\left(panjang\ \ 2(v), panjang\ \ (w)\right)} \qquad (1)$$

2.2 *Feature Selection*

Feature Selection adalah sebuah proses yang bisa digunakan pada *machine learning* di mana sekumpulan dari fitur yang dimiliki data digunakan untuk pembelajaran algoritma[8]. *Feature Selection* merupakan langkah penting dalam tahap *preprocessing*

yaitu untuk memilih atribut yang berpengaruh dan mengabaikan atribut yang tidak berpengaruh dalam proses pembuatan model atau klasifikasi *(Attribute Filtering)*.

2.3 *Missing Value*

Missing value adalah kondisi di mana terdapat data atau informasi yang tidak ditemukan pada suatu atribut tertentu dalam *dataset*[9]. *Missing value* dapat terjadi karena nilainya tidak relevan untuk kasus tertentu, tidak bisa dicatat pada saat data dikumpulkan, atau disebabkan adanya privasi[10].Untuk mengatasi *missing value*, dapat dilakukan beberapa hal seperti melakukan pengurangan objek data, memperkirakan nilai *missing values*, tidak melibatkan *missing values* dalam analisis data, dan mencari nilai rata-rata pada atribut yang memiliki *missing value*.

2.4 *Decision Tree*

Decision tree adalah bentuk sederhana teknik klasifikasi pada sekumpulan kelas tak berhingga yang direpresentasikan ke dalam bentuk simpul *(node)* dan rusuk *(edge)*[11]. Biasanya, *Decision Tree* dipilih untuk menyelesaikan masalah dengan *output* yang bernilai diskrit. Metode *Decision Tree* dipilih dalam penelitian ini karena dianggap memiliki solusi yang baik terhadap data yang memiliki *missing value*[12].

2.5 *Naive Bayes*

Naive Bayes mendasarkan pada asumsi penyederhanaan dimana nilai atribut secara kondisional saling bebas apabila diberikan nilai *output*[13]. Metode ini merupakan sebuah metode yang berakar pada teorema *Bayes*. Persamaan (2) merupakan persamaan Teorema *Bayes* yang menyatakan bahwa :

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \qquad (2)$$

Keterangan :

P(B|A)	=	probabilitas posterior, probabilitas muncul B jika diketahui A
P(B|A)	=	probabilitas posterior, probabilitas muncul A jika diketahui B
P(A)	=	probabilitas prior, probabilitas kejadian A
P(B)	=	probabilitas prior, probabilitas kejadian B

3 Metodologi Penelitian

3.1 *Dataset*

Data yang digunakan merupakan data pencatatan pendaftaran pasien kanker yang dikumpulkan dari tahun 2005 sampai dengan 2008 yang diperoleh dari https://archive.ics.uci.edu. Contoh nyata permasalahan yang ditimbulkan yaitu adanyanya kesamaan atribut pada pasien yang diduga orang tersebut adalah pasien yang sama. Hal ini dapat dilihat dari komponen atribut penyusunnya yaitu: *first name*,

last name, *sex*, *date of birth* (membandingkan kesamaan tanggal, bulan dan tahun) , serta *postal code*.

3.2 Pengecekan Duplikasi Data

Dalam penelitian ini, data yang digunakan berasal dari *dataset* pasien kanker yang dikumpulkan dari tahun 2005 sampai dengan 2008. Dari sekian data, *sample* yang digunakan dalam perhitungan sebanyak 2.266.941. Dari sekian banyak data yang digunakan, ada kemungkinan data dimiliki oleh beberapa pasien yang sama, sehingga perlu adanya pengecekan data untuk mengetahui apakah benar data tersebut dimiliki oleh pasien yang sama. Tabel 1 memperlihatkan *sample input* data yang diproses dalam penelitian ini.

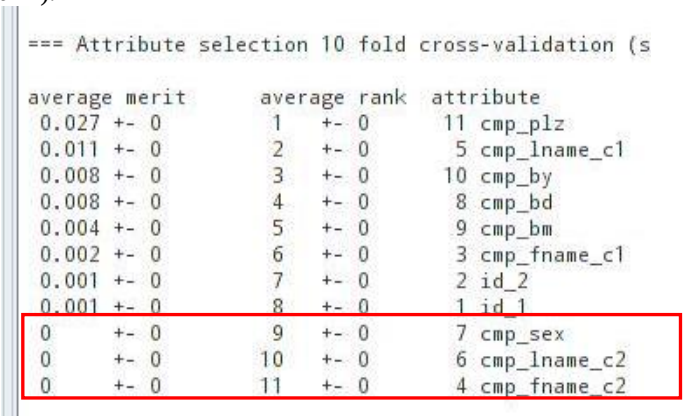
Tabel 1. *Sample* representasi data

'Alan',	'Dwi',	'Prasetyo',		'm',	'03',	'06',	'1994',	'55584',
'Alan',	,	'Prasetya',		'm',	'30',	'06',	'1994',	'55584',

Kedua data di atas kemudian digambarkan ke dalam bentuk kode biner (1,0,0,0,1,0,1,1,1). Setelah itu, diperoleh pola perbandingan dengan menggunakan *Levenshteain string metric* yaitu (1,0,0.875,0,1,0.5,1,1,1).

3.3 *Feature Selection*

Penelitian ini menggunakan *ranking selection* pada tahap *feature selection*. *Rangking selection* yaitu pemberian ranking pada setiap atribut yang ada dan mengabaikan *feature* yang tidak memenuhi standart tertentu. Proses *feature selection* yang dilakukan pada *tools* Weka dengan 10 *cross-validation* diperoleh 5 atribut terbawah yang tidak terlalu berpengaruh terhadap akurasi perhitungan. Dalam kasus ini, 5 atribut yang dibaikan, yaitu: id_1, id_2, cmp_sex, cmp_fname_2, cmp_lname_2 (Lihat Gambar 1).



```
=== Attribute selection 10 fold cross-validation (s

average merit      average rank  attribute
 0.027 +- 0          1   +- 0      11 cmp_plz
 0.011 +- 0          2   +- 0       5 cmp_lname_c1
 0.008 +- 0          3   +- 0      10 cmp_by
 0.008 +- 0          4   +- 0       8 cmp_bd
 0.004 +- 0          5   +- 0       9 cmp_bm
 0.002 +- 0          6   +- 0       3 cmp_fname_c1
 0.001 +- 0          7   +- 0       2 id_2
 0.001 +- 0          8   +- 0       1 id_1
 0     +- 0          9   +- 0       7 cmp_sex
 0     +- 0         10   +- 0       6 cmp_lname_c2
 0     +- 0         11   +- 0       4 cmp_fname_c2
```

Gambar 1. Penanganan *attribute selection*

3.4 Penanganan *Missing Value*

Data pasien kanker yang digunakan pada penelitian ini banyak sekali memuat *missing value*. Penanganan *missing value* pada penalitian ini dilakukan dengan *mean imputation*. *Mean imputation* adalah metode yang cukup sering digunakan. Metode ini mengganti *missing value* pada atribut dengan nilai rata-rata yang diperoleh dari seluruh atribut yang diketahui nilainya[14]. Tabel 2 menggambarkan jumlah *missing value* dan nilai rata-rata atribut.

Tabel 2. Atribut yang memuat *missing value*

Atribut	Jumlah missing values	Nilai *mean*
cmp_fname_c1	1007	0,713
cmp_bd	795	0,224
cmp_bm	795	0,489
cmp_by	795	0,223
cmp_plz	12843	0,006

3.5 Klasifikasi

Proses klasifikasi dilakukan setelah data selesai melewati tahap *preprocessing.* Pada tahapan *preprocessing* dilakukan pengecekan duplikasi data dengan *Levenshteain string metric*, pemilihan fitur, dan penanganan *missing value*. Data yang telah siap kemudian diklasifikasikan menggunakan metode *Decision Tree J48* dan *Naive Bayes* pada *tools* Weka.

3.6 Evaluasi dengan Tabel *Confusion Matrix*

Evaluasi hasil klasifikasi dilakukan dengan metode *confusion matrix*. *Confusion matrix* merupakan salah satu *tools* yang biasa digunakan dalam evaluasi mesin pembelajaran yang memuat dua atau lebih kategori[12,13]. *Confusion matriks* diperlihatkan pada Tabel 3.

Tabel 3. Tabel *Confusion Matrix* Dua Kelas

		Actual Class	
		Class-1	*Class-2*
Predicted Class	*Class-1*	*True positive*	*False negative*
	Class-2	*False positive*	*True negative*

3.7 Kurva *Receiver Operating Characteristic (ROC)*

Kurva ROC adalah salah satu teknik yang dapat memvisualisasikan, mengorganisasi, dan memilih *classifier* berdasarkan performanya.[15] *Receiver Operating Characteristic (ROC)* merupakan hasil dari pengukuran klasifikasi dalam bentuk 2 dimensi dimana garis horisontal menggambarkan nilai *false positive* dan garis vertikal sebagai *true positive*.[16]

Pada penelitian ini, tabel kontingensi yang digunakan untuk menganalisis *ROC* adalah yaitu tabel *Confusion Matrix* Dua Kelas.Visualisasi hasil perhitungan digambarkan dengan *Area Under ROC Curve* (AUC). AUC sering digunakan untuk mengukur kualitas *classifier* probabilistik.[15,17] Level pengukuran kualitas *classifier* menggunakan ROC dilihat berdasarkan akurasi dengan rentang yang diperlihatkan pada Tabel 4.[18]

Tabel 4. Nilai Kualitas *Classifier*

Rentang Akurasi	Kualitas *Classifier*
0.90-1.00	*Excellent*
0.80-0.90	*Good*
0.70-0.80	*Fair*
0.60-0.70	*Poor*
0.50-0.60	*Failure*

4 Hasil dan Pembahasan

Penelitian ini melakukan uji coba *dataset* pasien sebanyak 2.266.941 sebagai *sample* untuk dimodelkan. Proses pemodelan diawali dengan melakukan *feature selection*. Tahap *feature selection* dilakukan dengan dua tahap yaitu *attribute filtering* dan *ranking selection*. *Attribute filtering* mengabaikan atribut pada data yang tidak memiliki pengaruh dalam pencarian model. Atribut yang di-*filter* diperoleh dari proses *ranking selection* dimana lima buah atribut paling bawah yang tidak berpengaruh akan diabaikan dalam proses selanjutnya. Proses berikutnya adalah *replace missing values*. *Replace missing values* dilakukan dengan mengganti nilai yang hilang pada *record*. Nilai yang hilang digantikan dengan rata-rata seluruh data pada atribut tertentu.

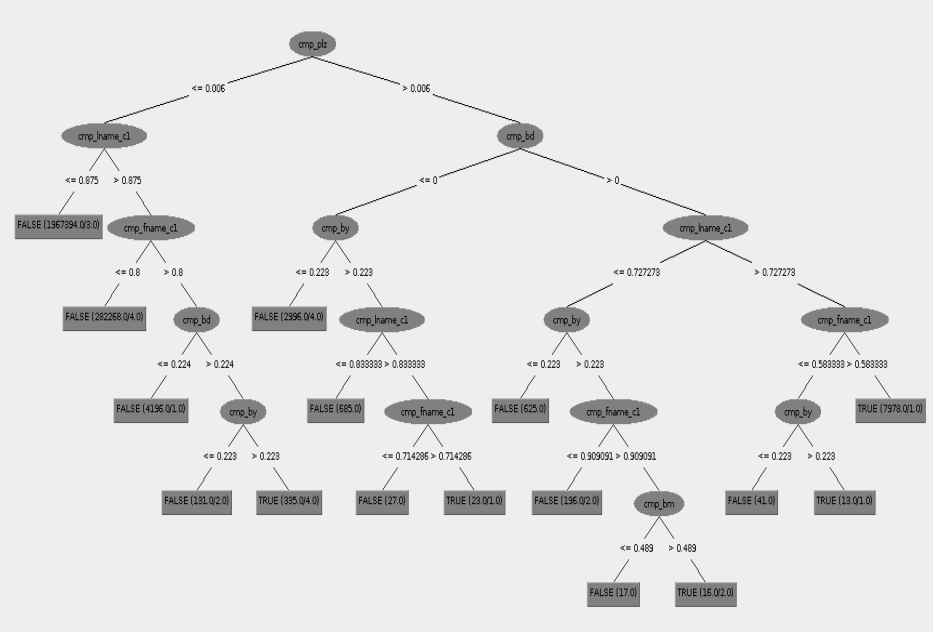
Data yang telah melalui tahap *replace missing values* kemudian diproses untuk selanjutnya dibuat model dan diklasifikasikan. Metode untuk membuat model data dan klasifikasi adalah *Decision Tree*. Selain itu, *Naive Bayes* digunakan sebagai pembanding metode *Decision Tree*.

Klasifikasi dilakukan dengan *tools* Weka menggunakan algoritma *Decision Tree J48* dan *Naive Bayes*. Pengujian hasil klasifikasi ditunjukkan dengan *cross validation* menggunakan 10 *fold*. Hasil *cross validation* dengan algoritma *Decision Tree J48* diperlihatkan oleh Tabel 5. Hasil akurasi memperlihatkan bahwa *Decision Tree J48* mencapai akurasi hingga 99,9988%.

Tabel 5. *Cross Validation* dengan *Decision Tree J48*

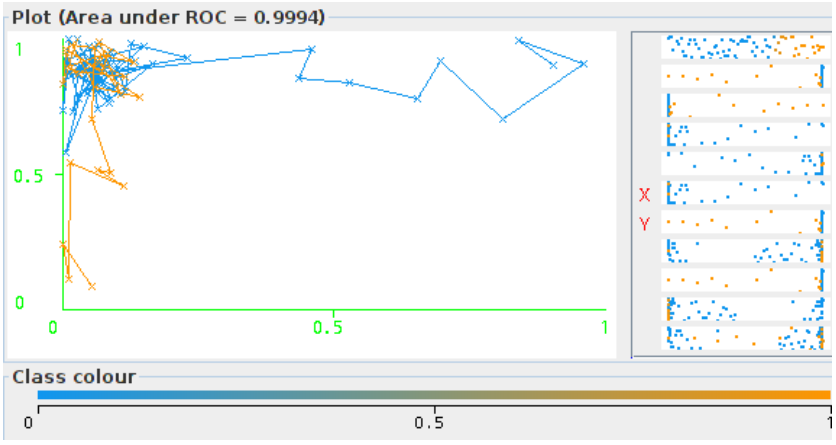
Akurasi: 99.9988 %		*Prediction Class*	
		True	*False*
Actual Class	*True*	8355	17
	False	11	2258558

Model *decision tree yang* terbentuk diperlihatkan oleh Gambar 2.

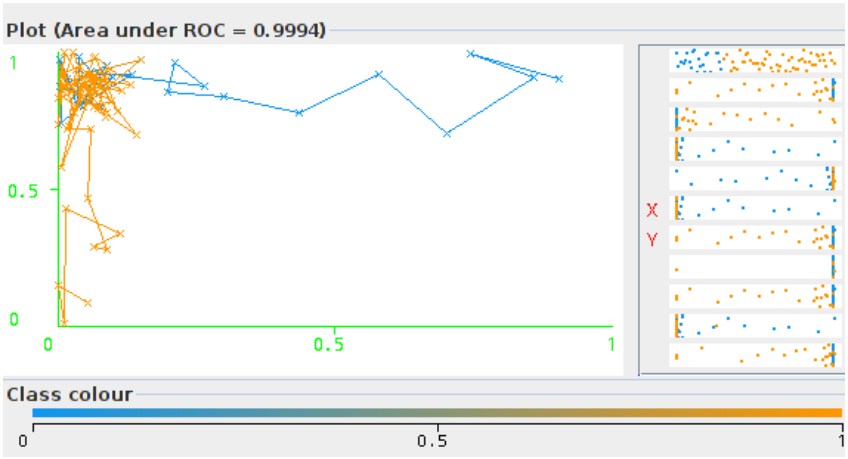


Gambar 2. Model *Decision Tree*

Visualisasi dengan *Decision Tree J48* diperlihatkan pada Gambar 3 dan Gambar 4 yang memperlihatkan visualisasi kurva ROC pada masing-masing atribut kelas 0 dan 1. Bagian atas grafik menampilkan sebuah nilai dari *Area Under ROC Curve* (AUC). AUC menjelaskan menunjukkan bahwa kinerja *Decision Tree classifier* memberikan AUC nilai sebesar 0,99994. Hal tersebut ditunjukkan dengan warna biru sebagai kelas 1 (*True)* dan warna *orange* sebagai kelas 0 *(False).*



Gambar 3. Kurva ROC Target Kelas 1 *(TRUE)*



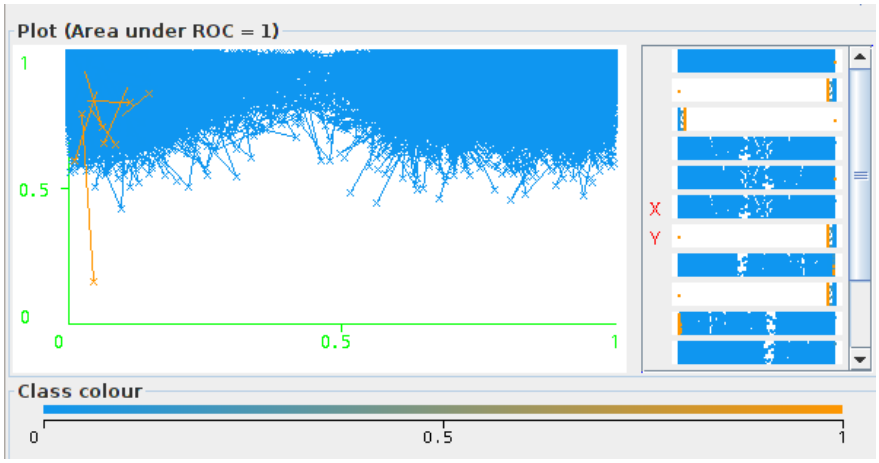
Gambar 4. Kurva ROC Target Kelas 0 *(FALSE)*

Hasil *cross validation* dengan *Naive Bayes* diperlihatkan oleh Tabel 6. Akurasi yang diperoleh dengan *Naive Bayes* yaitu sebesar 99.9799%.

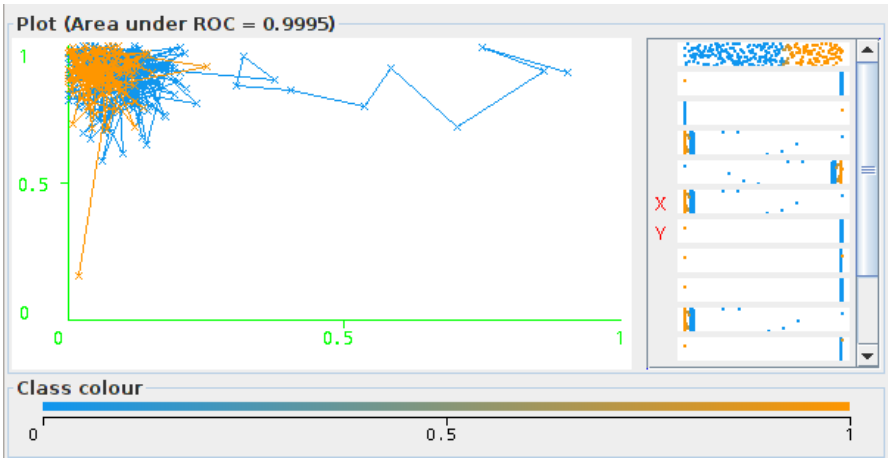
Tabel 6. *Cross Validation* dengan *Naive Bayes*

Akurasi: 99.9799 %		*Prediction Class*	
		True	*False*
Actual Class	*True*	8029	343
	False	113	2258456

Visualisasi performa *Naive Bayes* dengan ROC diperlihatkan oleh Gambar 5 dan Gambar 6. *Naive Bayes classifier* memberikan AUC nilai pada kelas 1 *(True)* sebesar 1 dan untuk kelas 0 *(False)* adalah 0,99995.



Gambar 5. Kurva ROC Target Kelas 1 *(TRUE)*



Gambar 6. Kurva ROC Target Kelas 0 *(FALSE)*

5 Kesimpulan

Penelitian ini telah menyelesaikan masalah *missing value* data pasien kanker dengan *replace missing value*. *Replace missing value* mengganti nilai yang hilang pada suatu atribut dengan nilai rata-rata suatu atribut. Untuk menentukan kemiripan data pasien, digunakan metode *Levenshtein string metric*. Metode *Levenshtein string metric* melakukan pengecekan setiap karakter string pada setiap atribut data pasien. Berdasarkan dua metode yang diuji coba, disimpulkan bahwa *Decision Tree* menghasilkan performa yang sedikit lebih baik dari *Naive Bayes* dalam menentukan klasifikasi. Metode *Decision tree* menghasilkan akurasi sebesar 99.9988 % sedangkan metode *Naive Bayes* menghasilkan akurasi 99.9799 %. Berdasarkan uji coba yang telah dilakukan, diperoleh bahwa *Decision Tree J48* memberikan hasil akurasi sedikit lebih baik dari *Naive Bayes* dengan selisih akurasi 0,0189%. Namun demikian, secara umum metode *Decision Tree* dan *Naive Bayes* sama-sama memiliki akurasi yang baik dalam melakukan klasifikasi kemiripan data pasien.

Pustaka

1. Jermyn, P., Dixon, M., & Read, B. J. (1999). Preparing Clean Views of Data For Data Mining. *ERCIM Work on Database Res*.
2. Kolcz, A., Chowdury, A., & Alspector, J. (2003). Data Duplication : An Imbalance Problem? *Workshop on Learning from Imbalanced Datasets II, ICML*.
3. Steorts, R. C., Hall, R., & Fienberg, S. E. (2014). SMERED: A Bayesian Approach to Graphical Record Linkage and De-duplication. *arXiv preprint arXiv:1403.0211*.

4. Sariyar, M., Borg, A., & Pommerening, K. (2012). Active Learning Strategies for The Deduplication of Electronic Patient Data Using Classification Trees. *Journal of Biomedical Informatics , 45* (5), 893-900.
5. Tamilselvi, J. J., & Gifta, C. B. (2011). Handling Duplicate Data in Data Warehouse for Data Mining. *International Journal of Computer Applications (0975 – 8887) , 15* (4).
6. Adiwidya, B. M. (2009). Algoritma Levenshtein Dalam Pendekatan Approximate String Matching. *Makalah IF3051 Strategi Algoritma*.
7. Haldar, R., & Mukhopadhyay, D. (2011), Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach, *arXiv:1101.1232*
8. Sewell, M. (2007). Feature Selection. *Available on http://machine-learning.martinsewell. com* .
9. Hermawati, F. A. (2009). *Data Mining.* Yogyakarta: Penerbit Andi.
10. Gimpy, Vohra, R., & Minakshi. (2014). Estimation of Missing Values Using Decision Tree Approach. *(IJCSIT) International Journal of Computer Science and Information Technologies , 5* (4), 5216-5220.
11. Santosa, B. (2007). *Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis.* Surabaya: Graha Ilmu.
12. Horn, C. (2010). *Analysis and Classificarton of Twitter Messages.* Master's Thesis, Graz University of Technology, Austria.
13. Manning, C., Raghavan, P., & Schutze, H. (2009). *Introduction to Information Retrieval.* Cambridge University Press.
14. Singh, S., & Prasad, J. (2013). Estimation of Missing Values in the Data Mining and Comparison of Imputation Methods. *Mathematical Journal of Interdisciplinary Sciences*, *1* (1), 75–90.
15. Vuk, M., & Curk, T. (2006). ROC Curve, Lift Chart And Calibration Plot. *Metodolo˘ski zvezki*, 3(1), 89-108.
16. Vercellis, C. (2009). *Business Intelligence: Data Mining and Optimization for Decision Making*. United Kingdom: John Willey & Son.
17. Eng, J. (2005). Receiver Operating Characteristic Analysis: A Primer1. *Academic radiology*, 12(7), 909-916.
18. Gorunescu, F. (2011). *Data Mining Concept Model and Techniques*. Berlin: Springer. ISBN 978-3-642-19720-8.