KOMPARASI METODE DATA MINING DALAM MEMPREDIKSI NASABAH BANK YANG AKAN MEMILIH TABUNGAN DEPOSITO MENGGUNAKAN ALGORITMA KLASIFIKASI

Wachid Darmawan
STMIK Widya Pratama Pekalongan
Jl. Patriot 25 Pekalongan Telp (0285) 427816
email: wachid.dw@gmail.com

Abstrak

Perkembangan trasaksi elektronik yang semakin pesat membuat banyak bank membuka unit-unit di tempat-tempat stategis untuk mengakomodir banyaknya nasabah. Dengan banyaknya nasabah bank juga perlu untuk mendapatkan nasabah yang bersedia untuk membuka tabungan deposito. Dengan banyaknya bank yang ada persaingan dalam mencari nasabah juga sangat beragam. Untuk membantu upacaya pihak marketing bank perlu mempelajari suatu ilmu yang bisa digunakan untuk memprediksinya nasabah yang bersedia membuka tabungan deposito, yaitu menggunakan algoritma klasifikasi yang ada di data mining. Komparasi dalam penelitian ini akan menggunakan algoritma klasifikasi, diantarnya: algoritma Decission Tree, algoritma K-Nearest Neighbour dan algoritma Naive Bayes untuk memprediksi nasabah yang akan membuka tabungan deposito. Dalam komparasi algoritma klasifikasi yang dilakukan pada penelitian ini dengan menggunakan software Rapid Miner 5.3, didapatkan hasil sebagai berikut: algoritma Decission Tree mendapatkan akurasi yang paling baik yaitu sebesar 88.48% , sedangkan algoritma Naive Bayes menghasilakan tingkat akurasi sebesar 86.84% dan algoritma K-Nearest Neighbour tingkat akurasinya sebesar 84.96%. Untuk penelitian kedepan diharapkan agar bisa dibuatkan software berbasis website supaya bisa digunakan oleh banyak orang.

Kata kunci: *klasifikasi deposito, Naive Bayes, K-Nearest Neighbour, Decission Tree*

1. Pendahuluan

1.1. Latar Belakang

Perkembangan transaksi elektronik didunia bisnis semakin hari semakin berkembang, dengan perkembangan tersebut banyak bank yang menawarkan investasi dengan risiko kecil, salah satunya dengan membuka tabungan deposito dibandingkan dengan investasi emas, saham atau lainya. Selain mempunyai risiko yang kecil deposito juga menarik nasabah karena tingkat suku bunganya yang relatif tinggi (Chi and Li 2017). Teknik marketing yang digunakan bank untuk mendapatkan nasabah juga berbeda-beda. Untuk itu pihak marketing saat akan melakukan klasifikasikan nasabahnya bisa menggunakan data nasabah lama untuk memprediksinya apakah nasabah tersebut bisa diajak kerjasama lagi apa tidak, bisa menggunakan ilmu data mining dengan melihat pola yang dihasilkan (Chi and Li 2017)(Apiletti et al. 2017).

Dengan perkembangan ilmu data mining yang semakin berkembang banyak penelitian yang digunakan untuk memprediksi suatu kasus perbankkan, seperti klasifikasi persetujuan kredit (Lobo 2017), nasabah kredit potensial (Chi and Li 2017) dan nasabah deposito potensial (Prabowo 2014). Dalam ilmu data mining salahsatunya adalah teknik klasifikasi. Teknik klasifikasi adalah suatu cara untuk menemukan pengetahuan baru dari data yang sudah lampau (Sugianti 2012). Banyak tenik klasifikasi yang digunakan dan mendapatkan tingkat akurasi yang baik atau tinggi.

Algoritma klasifikasi yang banyak digunakan dalam penelitian adalah algoritma Naive Bayes, algoritma K-Nearest Neighbour dan algoritma Decission Tree (Aggarwal 2015). Dalam data mining penggunakan tipe data akan berpengaruh dalam perhitungan algoritma yang digunakan (Gorunescu 2011), dengan kata lain algoritma a tidak bisa bekerja dengan baik pada tipe data a, atau sebaliknya tipe data a akan lebih baik pada lagoritma (Thakur and Juneja 2018). Penelitian ini akan membandingkan tinggkat akurasi yang dihasilkan antara algoritma *Naive Bayes*, algoritma *K-Nearest Neighbour* dan algoritma

Decission Tree untuk memprediksi nasabah yang akan membuka tabungan deposito. Pada tabel 1 memperlihatkan dataset bank marketing didapatkan dari repositori UCI.

Tabel 1. Data Nasabah (Moro et al., 2014)

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
48	management	married	tertiary	yes	-13	yes	no	cellular	15	may	20	6	291	2	other	no
37	admin.	married	tertiary	no	801	no	no	cellular	11	aug	331	7	-1	0	unknown	no
66	retired	married	tertiary	no	1948	no	no	cellular	28	jan	216	1	91	4	success	yes
32	management	single	tertiary	no	751	yes	no	unknown	8	may	64	1	-1	0	unknown	no
45	blue-collar	married	secondary	no	784	yes	yes	unknown	3	jun	34	1	-1	0	unknown	no
54	entrepreneur	married	unknown	no	1956	no	no	cellular	19	nov	221	5	-1	0	unknown	no
37	blue-collar	divorced	secondary	no	0	yes	no	cellular	21	apr	146	4	-1	0	unknown	yes
26	student	single	secondary	no	-147	no	no	unknown	4	jun	95	2	-1	0	unknown	no
34	technician	single	secondary	no	179	no	no	cellular	19	aug	294	3	-1	0	unknown	no
55	blue-collar	married	primary	no	1086	yes	no	cellular	6	may	146	1	272	2	failure	no
55	blue-collar	married	secondary	no	471	yes	no	unknown	30	may	58	2	-1	0	unknown	no
34	entrepreneur	married	tertiary	no	105	yes	no	unknown	28	may	249	2	-1	0	unknown	no
41	entrepreneur	divorced	secondary	no	1588	yes	yes	unknown	20	jun	10	8	-1	0	unknown	no
38	housemaid	divorced	secondary	no	70	no	no	cellular	27	oct	255	3	148	1	success	yes
43	technician	married	secondary	no	0	yes	yes	cellular	4	feb	539	8	204	5	failure	no
42	management	married	secondary	yes	-34	no	no	cellular	5	feb	176	1	-1	0	unknown	no
46	blue-collar	married	unknown	no	9328	yes	no	cellular	5	may	725	3	-1	0	unknown	yes
31	admin.	married	secondary	no	5	yes	yes	cellular	16	jul	101	7	-1	0	unknown	no
52	admin.	single	unknown	no	2227	no	no	cellular	24	feb	242	2	-1	0	unknown	yes
53	entrepreneur	married	primary	no	27	yes	no	telephone	10	sep	230	3	-1	0	unknown	yes
33	blue-collar	divorced	secondary	no	474	no	no	cellular	22	jul	400	2	-1	0	unknown	no
30	technician	single	tertiary	no	1185	yes	no	cellular	30	apr	284	1	339	1	failure	no
29	management	single	tertiary	no	1673	no	no	cellular	16	aug	200	2	159	3	other	no
46	services	married	secondary	no	2420	no	no	cellular	19	nov	405	2	-1	0	unknown	no
28	blue-collar	married	secondary	no	100	yes	yes	unknown	27	may	132	2	-1	0	unknown	no

2. Landasan Teori

2.1. Data Mining

Data mining salah satu cara untuk memunculkan pengetahuan yang ada di database atau *Knowledge Discovery in Database* (KDD) (Witten, Frank, and Hall 2011a)(Aggarwal 2015). Untuk mengetahui pola dan model suatu data perlu di lakukan ekstraksi data yang berguna untuk melihat informasi yang ada di dalamnya (Apiletti et al. 2017). Dengan adanya hubungan KDD dan data mining menyebabkan korelasi antar keduanya sangat dibutuhkan (Kavakiotis et al. 2017).

2.2. Klasifikasi

Dalam data mining teknik klasifikasi menempati salah satu peranan yang penting dalam proses mining. Dengan teknik klasifikasi bisa mentukan mana bagian dari data training dan hasil akurasinya yang di dapat dari data testing. Dalam penggunaanya teknik klasifikasi wajib mempunyai label didalam dataset (Falco-Walter, Scheffer, and Fisher 2018). Ada banyak algoritma yang digunakan dalam teknik klasifikasi, diantaranya: algoritma *Naive Bayes*, algoritma ID3, algoritma *K-Nearest Neighbour*, algoritma *Neural Network*, algoritma *Decission Tree* dan lain sebaginya (Aggarwal 2015).

2.3. Algoritma *Naive Bayes (NB)*

Untuk melakukan klasifikasi algoritma NB menggunakan teknik probabalitas, sehingga pada pengaplikasian didalam klasifikasi mempunyai tingkat akurasi yang cukup tinggi dengan kecepatan yang baik dalam mengolah dataset yang relatif besar (Fan 2009).

2.4. Algoritma *K-Nearest Neighbour (K-NN)*

Dalam melakukan klasifikasi algoritma *K-NN* dengan melakukan pencocokan suatu bobot dan fitur yang terdapat pada dataset. Penggunaan K pada algoritma K-NN digunakan untuk mentukan jumlah tetangga yang ada dan digunakan untuk mengambil keputusan (Ashari, Paryudi, and Tjoa 2013).

2.5. Algoritma *Decission Tree (C45)*

Algoritma yang simpel dan akurasinya baik salah satunya adalah *Decission Tree*. Algoritma C45 merupakan pengembangan dari algoritma CART dan ID3 (Ashari, Paryudi, and Tjoa 2013). Penggunaan algoritma C45 sangat populer dalam banyak penelitian karena hasil akhir dari algoritma C45 adalah suatu keputusan seperti pohon terbalik .

2.6. *Cross Validation*

Dalam mengolah dataset didalam teknik klasifikasi pada data mining banyak menggunakan teknik *cross validation*, karena akurasi yang dihasilkan sangat baik. *cross validation* merupakan suatu cara untuk memilih antara data testing dan data training dengan menggunakan komposisi tertentu. Biasanya menggunakan komposisi sepuluh bagian, sembilan bagian digunakan sebagai data training, sedangkan satu bagian digunakan sebagai data testing (Witten, Frank, and Hall 2011b).

2.7. *Confusion Matrix*

Penggunaan teknik *Confusion Matrix* biasanya digunakan dalam penelitian untuk mengevaluasi tingkat akurasi yang dihasilkan dari teknik klasifikasi pada proses mining. Penggunaan *Confusion Matrix* juga digunakan untuk mengetahu apakah klasifikasi tersebut bersifat positif atau negatif serta adakah kesalahan dalam melakukan proses klasifikasi (Gorunescu 2011).

3. Metode Penelitian

Untuk mengetahui komparasi yang ada dalam klasifikasi nasabah bank menggunakan metode eksperimen. Dalam penelitian ini menggunakan algoritma *Naive Bayes*, algoritma *K-Nearest Neighbour* dan algoritma *Decission Tree*.

3.1. Pengumpulan Data

Penggunaan dataset dalam penelitian ini adalah dataset dari repositorinya UCI, yaitu Bank Marketing dengan jumlah *record* sebanyak 4522. Jumlah atributnya adalah 17 *field* dengan 1 *field* untuk label, sedangkan 16 *field* lainya sebagai atribut.

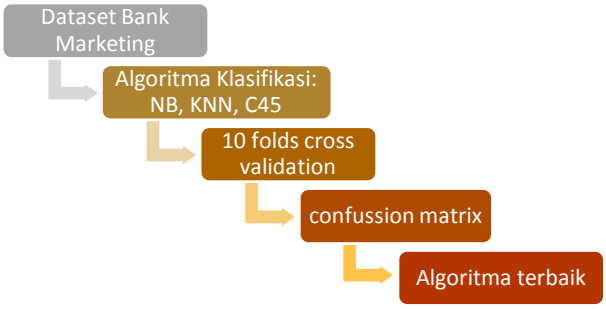
Tabel 2. Dataset Bank Marketing (Moro et al., 2014)

age	job	marital	education	default	balance	housing	loan	contact
48	management	married	tertiary	yes	-13	yes	no	cellular
37	admin.	married	tertiary	no	801	no	no	cellular
66	retired	married	tertiary	no	1948	no	no	cellular
32	management	single	tertiary	no	751	yes	no	unknown
45	blue-collar	married	secondary	no	784	yes	yes	unknown
54	entrepreneur	married	unknown	no	1956	no	no	cellular
37	blue-collar	divorced	secondary	no	0	yes	no	cellular
26	student	single	secondary	no	-147	no	no	unknown
34	technician	single	secondary	no	179	no	no	cellular
55	blue-collar	married	primary	no	1086	yes	no	cellular
55	blue-collar	married	secondary	no	471	yes	no	unknown
34	entrepreneur	married	tertiary	no	105	yes	no	unknown
41	entrepreneur	divorced	secondary	no	1588	yes	yes	unknown
38	housemaid	divorced	secondary	no	70	no	no	cellular
43	technician	married	secondary	no	0	yes	yes	cellular
42	management	married	secondary	yes	-34	no	no	cellular
46	blue-collar	married	unknown	no	9328	yes	no	cellular
31	admin.	married	secondary	no	5	yes	yes	cellular
52	admin.	single	unknown	no	2227	no	no	cellular
53	entrepreneur	married	primary	no	27	yes	no	telephone
33	blue-collar	divorced	secondary	no	474	no	no	cellular
30	technician	single	tertiary	no	1185	yes	no	cellular
29	management	single	tertiary	no	1673	no	no	cellular
46	services	married	secondary	no	2420	no	no	cellular
28	blue-collar	married	secondary	no	100	yes	yes	unknown

day	month	duration	campaign	pdays	previous	poutcome	y
15	may	20	6	291	2	other	no
11	aug	331	7	-1	0	unknown	no
28	jan	216	1	91	4	success	yes
8	may	64	1	-1	0	unknown	no
3	jun	34	1	-1	0	unknown	no
19	nov	221	5	-1	0	unknown	no
21	apr	146	4	-1	0	unknown	yes
4	jun	95	2	-1	0	unknown	no
19	aug	294	3	-1	0	unknown	no
6	may	146	1	272	2	failure	no
30	may	58	2	-1	0	unknown	no
28	may	249	2	-1	0	unknown	no
20	jun	10	8	-1	0	unknown	no
27	oct	255	3	148	1	success	yes
4	feb	539	8	204	5	failure	no
5	feb	176	1	-1	0	unknown	no
5	may	725	3	-1	0	unknown	yes
16	jul	101	7	-1	0	unknown	no
24	feb	242	2	-1	0	unknown	yes
10	sep	230	3	-1	0	unknown	yes
22	jul	400	2	-1	0	unknown	no
30	apr	284	1	339	1	failure	no
16	aug	200	2	159	3	other	no
19	nov	405	2	-1	0	unknown	no
27	may	132	2	-1	0	unknown	no

3.2. Kerangka Pemikiran

Untuk mengetahui hasil akhir dari penelitian ini akan menggunakan teknik *cross validation* dan untuk mengetahui tingkat akurasinya akan menggunakan teknik *confussion matrix*. Berikut kerangka pikir dalam penelitian ini, seperti pada gambar 1.



Gambar 1. Kerangka Pikir Penelitian

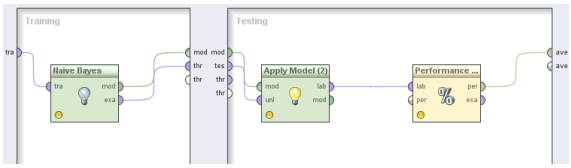
4. Hasil dan Pembahasan

4.1. Perhitungan Algoritma

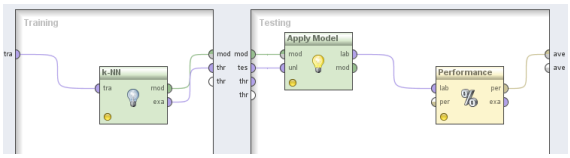
Dalam penelitian ini akan menggunakan software bantu Rapid Miner 5.3 untuk mengetahui komparasi dari algoritma NB, K-NN dan C.45. dalam proses perhitungan menggunakan algortima klasifikasi akan menggunakan teknik *cross validation* yang digunakan untuk melakukan perhitungan algoritma secara acak, meliputi satu bagian untuk data *testing* dan sembilan lainya untuk data *training*.

Untuk menghitung tingkat akurasi pada algoritma klasifikasi tersebut menggunakan tabel *confussion matrik*. Yang mana dalam percobaan yang sudah dilakukan dengan menggunakan software bantu Rapid Miner 5.3 sebanyak sepuluh kali, hasilnya kan di rerata untuk di lakukan komparasi dari algoritma klasifikasi dan menghasilkan akurasi terbaiknya.

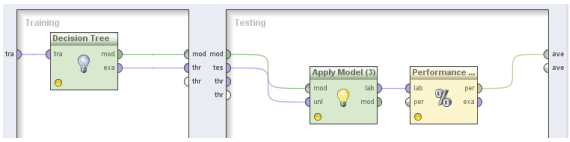
Untuk proses perhitungan yang dilakukan oleh software Rapid Miner 5.3, sebagaimana terlihat pada gambar 2, gambar 3 dan gambar 4.



Gambar 2. Perhitungan Algoritma NB



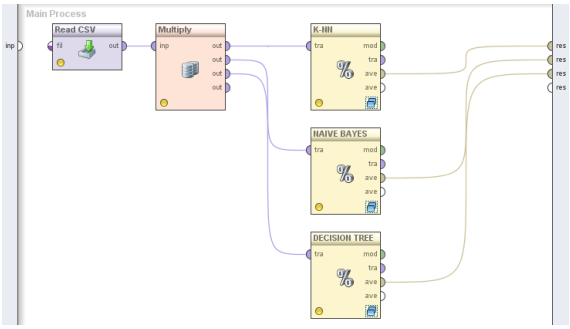
Gambar 3. Perhitungan Algoritma K-NN



Gambar 5. Perhitungan Algoritma K-NN

4.2. Hasil Komparasi

Dari hasil perhitungan yang dilakukan tiap algoritma klasifikasi selanjutnya digabungkan menjadi satu lembar kerja utama untuk mengetahui tingkat akurasi yang dihasilkan, seperti pada gambar 5.



Gambar 5. Lembar Kerja Komparasi Algoritma Klasifikasi

Setelah melakukan komparasi pada semua algoritma klasifikasi didapatkan hasil seperti yang terlihat di gambar 6, gambar 7 dan gambar 8 untuk masing-masing algoritma klasifikasi.

accuracy: 86.84% +/- 1.09% (mikro: 86.84%)

	true no	true yes	class precision
pred. no	3663	258	93.42%
pred. yes	337	263	43.83%
class recall	91.57%	50.48%	

Gambar 6. Hasil Akurasi Algoritma NB

accuracy: 84.96% +/- 0.94% (mikro: 84.96%)

	true no	true yes	class precision
pred. no	3678	358	91.13%
pred. yes	322	163	33.61%
class recall	91.95%	31.29%	

Gambar 7. Hasil Akurasi Algoritma K-NN

accuracy: 88.48% +/- 0.06% (mikro: 88.48%)

	true no	true yes	class precision
pred. no	4000	521	88.48%
pred. yes	0	0	0.00%
class recall	100.00%	0.00%	

Gambar 8. Hasil Akurasi Algoritma C.45

4.3. Pembahasan

Setelah mengetahui hasil komparasi yang dihasilkan oleh masing-masing algoritma klasifikasi dengan menggunakan software bantu Rapid Miner 5.3 seperti tertera pada gambar 6, gambar 7 dan gambar 8. Langkah selanjutnya adalah melakukan komparasi dari akurasi masing-masing algoritma NB, K-NN dan C.45 menggunakan tabel *confussion matrix*, yang ada pada tabel 3.

Tabel 3. Hasil Algoritma Komparasi

Algoritma	Akurasi	Micro
NB	86.84	86.84
K-NN	84.96	84.96
C.45	88.48	88.48

5. Kesimpulan dan Saran

5.1. Kesimpulan

Dataset yang digunakan sudah banyak digunakan oleh peneliti lain untuk membuat pemodelan dan untuk membuat aplikasi klasifikasi, yaitu dataset Bank Marketing dari repositori UCI.

Dari hasil akurasi yang sudah dihitung menggunakan software bantu Rapid Miner 5.3 didapat hasil bahwa algoritma *Decission Tree* mendapatkan akurasi yang paling baik yaitu sebesar 88.48% , sedangkan algoritma *Naive Bayes* menghasilakan tingkat akurasi sebesar 86.84% dan algoritma *K-Nearest Neighbour* tingkat akurasinya sebesar 84.96%.

5.2. Saran

Untuk penelitian ini masih menggunakan software bantu dalam menghitung tingkat akurasi yang dihasilkan oleh masing-masing algoritma NB, K-NN dan C.45. Untuk penelitian kedepan diharapkan agar bisa dibuatkan software berbasis website supaya bisa digunakan dan diakses oleh banyak orang.

6. Daftar Pustaka

Aggarwal, Charu C. 2015. *Data Classification : Algorithms and Applications. Series: Chapman & Hall/CRC Data Mining and Knowledge Discovery Series ; 35*.

Apiletti, Daniele, Elena Baralis, Tania Cerquitelli, Paolo Garza, Fabio Pulvirenti, and Luca Venturini. 2017. "Frequent Itemsets Mining for Big Data: A Comparative Analysis." *Big Data Research* 9. Elsevier Inc.: 67–83.

Ashari, Ahmad, Iman Paryudi, and A Min Tjoa. 2013. "Performance Comparison between Naïve Bayes , Decision Tree and K-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool" 4 (11): 33–39.

Chi, Qinwei, and Wenjing Li. 2017. "Economic Policy Uncertainty, Credit Risks and Banks' Lending Decisions: Evidence from Chinese Commercial Banks." *China Journal of Accounting Research* 10 (1). Sun Yat-sen University: 33–50.

Falco-Walter, Jessica J., Ingrid E. Scheffer, and Robert S. Fisher. 2018. "The New Definition and Classification of Seizures and Epilepsy." *Epilepsy Research* 139 (November 2017): 73–79.

Fan, Liwei. 2009. "Improving the Naïve Bayes Classifier," 879–83.

Gorunescu, Florin. 2011. *Data Mining: Concepts and Techniques*. *Chemistry &*

Kavakiotis, Ioannis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, and Ioanna Chouvarda. 2017. "Machine Learning and Data Mining Methods in Diabetes Research." *Computational and Structural Biotechnology Journal* 15. The Authors: 104–16.

Lobo, Gerald J. 2017. "Accounting Research in Banking – A Review." *China Journal of Accounting Research* 10 (1). Sun Yat-sen University: 1–7.

Prabowo, Alvino Dwi Rachman. 2014. "Prediksi Nasabah Yang Berpotensi Membuka Simpanan Deposito Menggunakan Naive Bayes Berbasis Particle Swarm Optimization." *Fakultas Ilmu Komputer, Universitas Dian Nuswantoro Semarang*, no. 5.

Sugianti, Devi. 2012. "Algoritma Bayesian Classification Untuk Memprediksi Heregistrasi Mahasiswa Baru Di STMIK Widya Pratama," no. 2: 1–5.

Thakur, Niharika, and Mamta Juneja. 2018. "Survey on Segmentation and Classification Approaches of Optic Cup and Optic Disc for Diagnosis of

Glaucoma." *Biomedical Signal Processing and Control* 42. Elsevier Ltd: 162–89.
Witten, Ian H., Eibe Frank, and Mark a. Hall. 2011a. *Data Mining*. *Data Mining*. Vol. 277.
———. 2011b. *Data Mining: Practical Machine Learning Tools and Techniques*. Third Edit. USA: Morgan Kaufmann Publishers.