Teknik Machine Learning untuk Bank Marketing Dataset

Fahmi Izhari
Universitas Pembangunan Pancabudi Medan
Email : fahmi_izhari@dosen.pancabudi.ac.id

Abstrak

Peneliti melakukan penelitian ini agar dapat memprediksi manakah klien yang telah berlangganan deposito berjangka dengan baik atau macet agar dapat mempermudah pihak bank dalam mengetahui keakurasian data dengan melihat tingkat akurasi dan tingkat error berdasarkan Coinfussion Matrix. Pada penelitian ini, untuk mengetahui kinerja dari Algoritma NN, SVM, Decision Tree, Rules Based, Fuzzy Logic, Regression Model, KNN, Random Forest, Naïve Bayes maka digunakan data observasi berupa dataset UCI Bank Marketing yang berasal dari (http://archive.ics.uci.edu/ml). Data Bank Marketing ini melibatkan 45212 sampel dengan 16 faktor serta 2 label kelas meliputi: apakah klien telah berlangganan deposito berjangka? ('yes', 'no'). penelitian ini mendapatkan hasil akurasi dengan algoritma NN sebesar 87.54%, SVM sebesar 73.45%, Decision Tree sebesar 75.64%, Rules Based sebesar 65.32%, Fuzzy Logic sebesar 71.34%, Regression Model sebesar 82.32%, KNN sebesar 62.54%, Random Forest sebesar 68.77%, Naïve Bayes sebesar 83.78%.

***Kata Kunci** : Bank Marketing, Algoritma NN, SVM, Decision Tree, Rules Based, Fuzzy Logic, Regression Model, KNN, Random Forest, Naïve Bayes.*

Abstract

Researchers conducted this research in order to predict which clients had subscribed to time deposits well or were stuck in order to make it easier for the bank to determine the accuracy of the data by looking at the level of accuracy and error level based on the Coinfusion Matrix. In this study, to determine the performance of the NN Algorithm, SVM, Decision Tree, Rules Based, Fuzzy Logic, Regression Model, KNN, Random Forest, Naïve Bayes, observational data is used in the form of the UCI Bank Marketing dataset originating from (http: // archive .ics.uci.edu / ml). This Bank Marketing data involves 45212 samples with 16 factors and 2 class labels including: has the client subscribed to a time deposit? ('yes', 'no'). This research got the results of accuracy with the NN algorithm of 87.54%, SVM of 73.45%, Decision Tree of 75.64%, Rules Based of 65.32%, Fuzzy Logic of 71.34%, Regression Model of 82.32%, KNN of 62.54%, Random Forest of 68.77 %, Naïve Bayes at 83.78%.

***Keywords**: Bank Marketing, NN Algorithm, SVM, Decision Tree, Rules Based, Fuzzy Logic, Regression Model, KNN, Random Forest, Naïve Bayes.*

1. Pendahuluan

Pada dunia perbankan perlu adanya sebuah teknik marketing, data marketing, dan lain-lain. Bank merupakan salah satu jenis lembaga keuangan yang menyediakan berbagai macam jasa produk bank yang diantaranya merupakan deposito, kredit, marketing, dan lain-lain. Tujuan klasifikasi adalah untuk memprediksi apakah klien akan berlangganan deposito berjangka (variabel y). Data tersebut nantinya akan terkait dengan kampanye pemasaran langsung dari lembaga perbankan Portugis. Kampanye pemasaran didasarkan pada panggilan telepon. Seringkali, lebih dari satu

kontak ke klien yang sama diperlukan, untuk mengakses apakah produk (deposito berjangka bank) akan ('ya') atau tidak ('tidak') berlangganan.

Klasifikasi yang dilakukan kepada objek yang didasari terhadap data pembelajaran yang memiliki jarak yang dekat dengan objek yang telah diuji. Pendekatan yang dilakukan pada pencarian kasus perhitungan pendekatan antara masalah baru dengan masalah sebelumnya dengan melakukan penyetaraan bobot dari penjumlahan fitur yangada, dalam Penanganan Bank Marketing tersebut dalam melakukan penerapan dari metode yang digunakan sangat dibutuhkan untuk proses pengolahan data yang baik untuk menganalisa kinerja algoritma melalui penerapan dari algoritma dalam penentuan klasifikasi pada data Nasabah

Pembelajaran adalah prosedur mengembangkan model setelah pengetahuan yang tercakup dari sebuah data, sedangkan pembelajaran mesin adalah hal yang kompleks prosedur komputasi untuk mengenali pola secara otomatis dan pengambilan keputusan yang baik berdasarkan sampel data yang terlatih. Machine learning memiliki kemampuan dalam bidang memprediksi, mengelompokkan, atau mengklasifikasikan tetapi tidak terlihat atau data baru berdasarkan pembelajaran atau pelatihannya. Beberapa teknik pembelajaran mesin yang terkenal termasuk Jaringan Syaraf Tiruan (JST), SVM, Decision Tree, Naïve Bayes dan kmean pengelompokan dll. Ada pendekatan ansambel juga yang mengintegrasikan hasil klasifikasi individu teknik dan menghasilkan kinerja yang lebih baik secara keseluruhan.

Beberapa peneliti/periset sering menemukan data dengan kondisi data dengan kelas tidak seimbang. Data dengan kelas tidak seimbang biasanya memiliki permasalahan didalam melakukan klasifikasi pada *machine learning*, karena jumlah data per kelas tidak terdistribusi secara merata/normal. Kondisi ini biasanya ditemukan pada data data kredit, kesehatan dan lainnya (Nikitin, 2018). Liu (2009) menyatakan bahwa algoritma pembelajaran yang tidak mempertimbangkan ketidakseimbangan pada kelas mayoritas cenderung kewalahan oleh kelas minoritas dalam melakukan prediksi.

Peneliti melakukan penelitian ini agar dapat memprediksi manakah klien yang telah berlangganan deposito berjangka dengan baik atau macet agar dapat mempermudah pihak bank dalam mengetahui keakurasian data dengan melihat tingkat akurasi dan tingkat error berdasarkan Coinfussion Matrix.

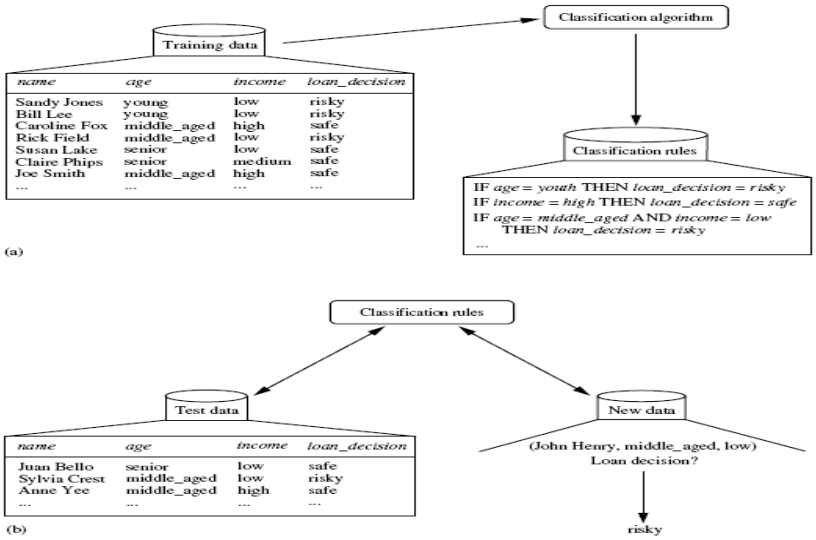
2. Metode Penelitian

Proses dalam menganalisa sebuah data dalam menemukan model yang dapat menguraikan atau mengelompokkan data-data kelas yang penting yang digunakan untuk memprediksi kelas dari objek yang tidak diketahui kelasnya merupakan proses yang dilakukan sebagai teknik klasifikasi. Sehingga modelnya ditemukan dengn cara analisa data yang diuji atau yang terdapat pada objek data yang diketahui kelasnya (Han, 2012). Model – model tersebut berupa algoritma klasifikasi yang pada umumnya sering digunakan dalam menganalisa data yang termasuk K-NN, *Genetic Algorithm, Rule Base,* C4.5, Naïve Bayesian dan lain sebagainya.

Teknik klasifikasi didasarkan pada empat komponen utama yaitu:

1. *Class label attribute.*
   Dalam proses ini melakukan pengkategorian untuk mempresentasikan label yang ada pada objek data. Misalnya: resiko penyakit, kresit maupun jenis pinjaman, dan lain sebagainya.
2. *Predictor*
   Variable ini mempresentasikan karakteristik pada atribut data. Misalnya: merokok atau tidak, pergi kesekolah atau tidak, dan lain sebagainya.
3. *Training Dataset*
   Sebuah dataset yang bernilai dari kelas komponen dan *predictor* digunakan dalam penentuan kelas yang cocok yang dasarnya yaitu *predictor*.
4. *Testing Dataset*
   Dalam pengklasifikasian pada model predictor menggunakan data baru yang hasilnya dari pengukuran akurasi klasifikasi dari metode evaluasi.

Pengklasifikasian data dapat diilustrasikan pada Gambar 1 berikut ini.



Gambar 1. Ilustrasi Kalsifikasi

Pada Gambar 1. diatas menjelaskan proses pembelajaran dengan data uji yang dianalisa dengan menerapkan algoritma klasifikasi. Atribut keputusan menjadi sebuah label kelas dan model klasifikasi dipresentasikan dalam bentuk rule klasifikasi. Sedangkan proses klasifikasi selanjutnya yang dipakai dalam pengestimasi keakurasian dari aturan klasifikasi yang dapat dihasilkan. Jika akurasi yang dihasilkan maka aturan dapat diperoleh dari data yang baru. (Han, et al. 2012).

Pada penelitian ini, untuk mengetahui kinerja dari Algoritma NN, SVM, Decision Tree, Rules Based, Fuzzy Logic, Regression Model, KNN, Random Forest, Naïve Bayes maka digunakan data observasi berupa dataset UCI *Bank Marketing* yang berasal dari (http://archive.ics.uci.edu/ml). Data Bank Marketing ini melibatkan 45212 sampel dengan 16 faktor serta 2 label kelas meliputi: *apakah klien telah berlangganan deposito berjangka? ('yes', 'no')*.

Adapun deskripsi dari dataset UCI *Bank Marketing* dapat dilihat pada tabel 1 sebagai berikut:

Tabel 1. *Dataset UCI Bank Marketing*

No.	age	job	marital	education	duration	…	y
1	58	management	married	tertiary	261	…	no
2	44	technician	single	secondary	151	…	no
3	33	entrepreneur	married	secondary	76	…	no
4	47	blue-collar	married	unknown	92	…	no
5	33	unknown	single	unknown	198	…	no
6	35	management	married	tertiary	139	…	no
7	28	management	single	tertiary	217	…	no
8	42	entrepreneur	divorced	tertiary	380	…	no
9	58	retired	married	primary	50	…	no
10	43	technician	single	secondary	55	…	no
11	41	admin.	divorced	secondary	222	…	no
12	29	admin.	single	secondary	137	…	no
13	53	technician	married	secondary	517	…	no
14	58	technician	married	unknown	71	…	no
15	57	services	married	secondary	174	…	no
16	51	retired	married	primary	353	…	no
17	45	admin.	single	unknown	98	…	no
18	57	blue-collar	married	primary	38	…	no
19	60	retired	married	primary	219	…	no
20	33	services	married	secondary	54	…	no
21	28	blue-collar	married	secondary	262	…	no
…	…	…	…	…	…	…	…

45212	37	entrepreneur	married	secondary	361	…	no

Berikut merupakan informasi atributnya:

1. usia
2. pekerjaan: jenis pekerjaan (kategori: 'admin.', 'Kerah biru', 'pengusaha', 'pembantu rumah tangga', 'manajemen', 'pensiunan', 'wiraswasta', 'layanan', 'pelajar' , 'teknisi', 'pengangguran', 'tidak diketahui'
3. perkawinan: status perkawinan (kategorikal: 'bercerai', 'menikah', 'lajang', 'tidak diketahui'; catatan: 'cerai' berarti bercerai atau janda)
4. pendidikan (kategorikal: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5. default: memiliki kredit dalam default? (kategorikal: 'tidak', 'ya', 'tidak diketahui')
6. perumahan: memiliki pinjaman perumahan? (kategorikal: 'tidak', 'ya', 'tidak diketahui')
7. pinjaman: memiliki pinjaman pribadi? (kategorikal: 'tidak', 'ya', 'tidak diketahui')
8. kontak: jenis komunikasi kontak (kategorikal: 'seluler', 'telepon')\
9. bulan: kontak terakhir bulan dalam setahun (kategorikal: 'jan', 'feb', 'mar', ..., 'nov', 'des')
10. day_of_week: hari kontak terakhir dalam seminggu (kategorikal: 'mon', 'tue', 'wed', 'thu', 'fri')
11. durasi: durasi kontak terakhir, dalam detik (numerik). Catatan penting: atribut ini sangat mempengaruhi target keluaran (misalnya, jika durasi = 0 maka y = 'tidak'). Namun, durasinya tidak diketahui sebelum panggilan dilakukan. Juga, setelah panggilan berakhir, y jelas diketahui. Dengan demikian, input ini hanya boleh dimasukkan untuk tujuan benchmark dan harus dibuang jika tujuannya adalah untuk memiliki model prediksi yang realistis.
12. kampanye: jumlah kontak yang dilakukan selama kampanye ini dan untuk klien ini (numerik, termasuk kontak terakhir)
13. hari: jumlah hari yang berlalu setelah klien terakhir dihubungi dari kampanye sebelumnya (numerik; 999 berarti klien sebelumnya tidak dihubungi)
14. sebelumnya: jumlah kontak yang dilakukan sebelum kampanye ini dan untuk klien ini (numerik)

3. Hasil dan Pembahasan

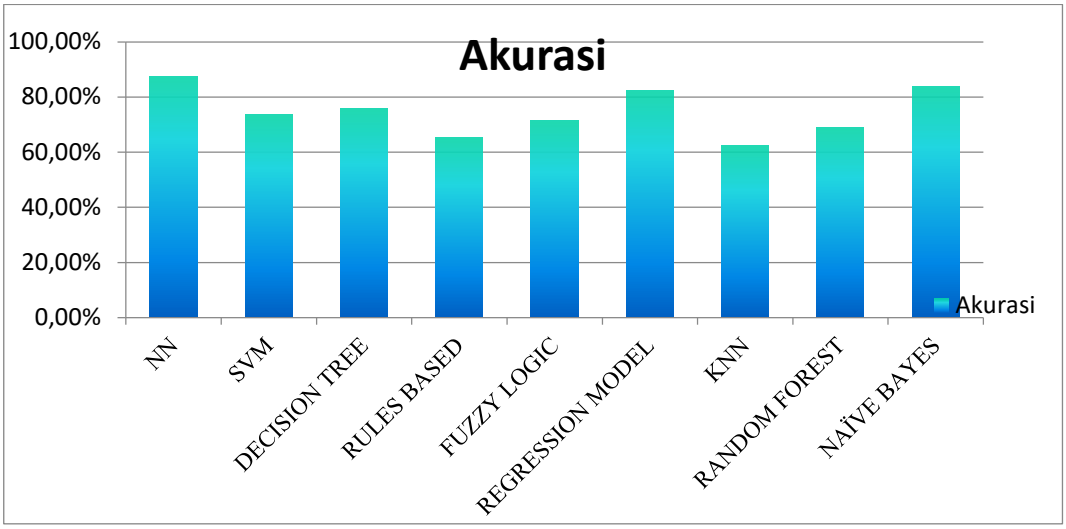
Pada penelitian ini, untuk mengetahui kinerja dari Algoritma NN, SVM, Decision Tree, Rules Based, Fuzzy Logic, Regression Model, KNN, Random Forest, Naïve Bayes maka digunakan data observasi berupa dataset UCI *Bank Marketing* yang berasal dari (http://archive.ics.uci.edu/ml). Data Bank Marketing ini melibatkan 45212 sampel dengan 16 faktor serta 2 label kelas meliputi: *apakah klien telah berlangganan deposito berjangka? ('yes', 'no').*

Adapun hasil analisa yang telah dilakukan adalah sebagai berikut:

Tabel 2. Hasil Akurasi

Algoritma	Akurasi
NN	87.54%
SVM	73.45%
DECISION TREE	75.64%
RULES BASED	65.32%
FUZZY LOGIC	71.34%
REGRESSION MODEL	82.32%
KNN	62.54%
RANDOM FOREST	68.77%
NAÏVE BAYES	83.78%

Berikut diagram akurasi dari penelitian:



Gambar 2. Tabel Akurasi

4. Kesimpulan

Penelitian ini telah dilakukan secara sistematis, untuk membantu pembaca mendapatkan pengetahuan sebelumnya penelitian yang dilakukan dalam menganalisa dataset Bank Marketing. Penelitian ini dilakukan untuk menganalisis sebuah data Bank Marketing agar dapat memprediksi manakah klien yang telah berlangganan deposito berjangka dengan baik atau macet agar dapat mempermudah pihak bank dalam mengetahui keakurasian data dengan melihat tingkat akurasi dan tingkat error berdasarkan Coinfussion Matrix.

Daftar Pustaka

[1] Dai Qin-yun,. Zang Chun-Ping., Wu Hao. 2016. *Research of Decision tree Classification Algorithm in Data Mining*. Dept. of Electric and Electronic Engineering, Shijiazhuang Vocational and Technology Institute. China

[2] Kotu, V. & Deshpande, B. 2015. *Predictive Analytics and Data Mining*. Morgan Kaufmann Publisher: San Francisco.

[3] Sahu, Mridu., Nagwani. N.K., Verma Shrish., Shirke. Saransh. 2015. *Performance Evaluation of Different Classifier for Eye State Prediction Using EEG Signal. International Journal of Knowledge Engineering*, Volume.1, No.2.

[4] Sharma, R., Purushottam, Saxena, K. 2016. Efficient Heart Disease Prediction System using Decision Tree. International Conference on Computing, Communication and Automation (ICCCA), Noida, India, 15-16 May. 72-77. DOI: 10.1109/CCAA.2015.7148346

[5] Sivapriya, T. R., Nadira, B. K. 2013.Hybrid Feature Selection for Enhanced Classification of High Dimensional Medical Data.*International Conference on Computational Intelligence and Computing Research*, pp. 1-4.