Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik20185958
Vol. 5, No. 5, Oktober 2018, hlm. 567-576 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 51/E/KPT/2017 e-ISSN: 2528-6579

PERBANDINGAN TEKNIK KLASIFIKASI DALAM *DATA MINING* UNTUK *BANK DIRECT MARKETING*

Irvi Oktanisa[1], Ahmad Afif Supianto[2]

[1,2] Program Studi Magister Ilmu Komputer, Fakultas Ilmu Komputer, Universitas Brawijaya
Email: [1]irvioktanisa@student.ub.ac.id, [2]afif.supianto@ub.ac.id

(Naskah masuk 25 Juli 2018, diterima untuk diterbitkan 30 Oktober 2018)

Abstrak

Klasifikasi merupakan teknik dalam *data mining* untuk mengelompokkan data berdasarkan keterikatan data terhadap data sampel. Pada penelitian ini, kami melakukan perbandingan 9 teknik klasifikasi untuk mengklasifikasi respon pelanggan pada *dataset Bank Direct Marketing*. Perbandingan teknik klasifikasi ini dilakukan untuk mengetahui model dalam teknik klasfikasi yang paling efektif untuk mengklasifikasi target pada *dataset Bank Direct Marketing*. Teknik klasifikasi yang digunakan yaitu *Support Vector Machine*, *AdaBoost*, *Naïve Bayes*, *Constant*, *KNN*, *Tree*, *Random Forest*, *Stochastic Gradient Descent*, dan *CN2 Rule*. Proses klasifikasi diawali dengan *preprocessing* data untuk melakukan penghilangan *missing value* dan pemilihan fitur pada *dataset*. Pada tahap evaluasi digunakan teknik *10 fold cross validation*. Setelah dilakukan pengujian, didapatkan bahwa hasil klasifikasi menunjukkan akurasi terbaik diperoleh oleh model *Tree*, *Constant*, *Naive Bayes*, dan *Stochastic Gardient Descent*. Kemudian diikuti oleh model *Random Forest*, *K-Nearest Neighbor*, *CN-2 Rule*, *AdaBoost* dan *Support Vector Machine*. Dari keempat model yang menunjukkan hasil akurasi terbaik, untuk kasus ini *Stochastic Gradient Descent* terpilih sebagai model yang memiliki akurasi terbaik dengan nilai akurasi sebesar 0,972 dan hasil visualisasi yang dihasilkan lebih jelas untuk mengklasifikasi target pada *dataset Bank Direct Marketing*.

Kata kunci: *Pebandingan, klasifikasi, data mining, decission tree, machine learning, bank direct marketing*

A COMPARISON OF CLASSIFICATION TECHNIQUES IN DATA MINING FOR BANK DIRECT MARKETING

Abstract

Classification is a technique in data mining to classify data based on the attachment of data to the sample data.. In this paper, we present the comparison of 9 classification techniques performed to classify customer response on the dataset of Bank Direct Marketing. The techniques performed to find out the effectiveness model in the classification technique used to classify targets on the dataset of Bank Direct Marketing. The techniques used are Support Vector Machine, AdaBoost, Naïve Bayes, Constant, KNN, Tree, Random Forest, Stochastic Gradient Descent, and CN2 Rule. The classification process begins with preprocessing data to perform missing value omissions and feature selection on the dataset. Cross validation technique, with k value is 10, used in the evaluation stage. After testing, it was found that the classification results showed the best accuracy obtained when using the Tree model, Constant, Naive Bayes and Stochastic Gradient Descent. Afterwards the Random Forest model, K-Nearest Neighbor, CN-2 Rule, AdaBoost, and Support Vector Machine are followed. Of the four models with the high accuracy results, in this case Stochastic Gradient Descent was selected as the best accuracy model with an accuracy value of 0.972 and resulting visualization more clearly to classify targets on the dataset of Bank Direct Marketing.

Keywords: *Comparison of classification, data mining, decission tree, machine learning, bank direct marketing*

1. PENDAHULUAN

Data mining merupakan proses untuk memanipulasi data dengan mengekstraksi informasi yang sebelumnya tidak diketahui dari *dataset* yang berukuran besar (Vijayakumar, & Nedunchezhian, 2012). Belakangan ini, *data mining* sering digunakan pada beberapa industri termasuk asuransi dan perbankan. Penggunaan teknik *data mining* dalam *Bank Direct Marketing* bertujuan untuk menganalisa data pelanggan dan mengembangkan data pelanggan secara statistika berdasarkan produk dan pelayanan yang lebih disukai oleh pelanggan.

Masalah yang dihadapi dalam *Bank Direct Marketing* ini adalah bagaimana mencapai akurasi yang tinggi dalam proses klasifikasi berdasarkan ketepatan informasi tertentu yang diperoleh dari *customer* dan dianggap penting oleh pihak bank (Elsalamony, 2013). Target utama pada *Bank Direct Marketing campaign* yaitu mencoba memprediksi harapan terhadap konsumen yang memiliki kemungkinan tertinggi dalam pelayanan menggunakan teknik *data mining* (Vaidehi, 2016). Implementasi dalam *Bank Direct Marketing* digunakan pada nasabah bank kredit. Bank harus selektif dalam memilih nasabah yang menerima kredit (Anggodo, dkk., 2017).

Dalam menyelesaikan permasalahan klasifikasi, penggunaan metode atau teknik bertujuan untuk mempermudah proses klasifikasi. Beberapa teknik yang digunakan dalam kasus klasifikasi yaitu *decission tree, clasification and assosiation rule*, *six-sigma methodology*, dan *CRISP methodology*. Penelitian terkait klasifikasi pernah dilakukan oleh Niu (2009) menggunakan *compactness of rule*, dengan hasil penelitian dengan metode yang diusulkan memiliki hasil klasifikasi terbaik dalam perbandingannya dengan teknik klasifikasi dan asosiasi berdasar pada *rule mining*. Bartik (2009) pada penelitiannya mengusulkan teknik klasifikasi *assosiation rule mining* untuk data relational dan *web mining*. Hao (2009) menggunakan *association rule* sebagai peningkatan metode klasifikasi pada klasifikasi data *Bank Direct Marketing*. Grzonka (2009) pada penelitiannya menggunakan *decission tree* untuk klasifikasi menggunakan pendekatan yang mendefenisikan skenario dimana pelanggan dari bank membuat keputasan tentang pengaktifan deposit mereka.

Penelitian terdahulu pada area *bank direct marketing* tentang klasifikasi telah dilakukan oleh Penelitian yang dilakukan oleh Elsalamony & Elsayad (2013) menggunakan teknik *data mining* yang di *hybri*d dengan *Multi Layer Perceptron*. Hasil menunjukkan, metode yang diusulkan memiliki akurasi yang tinggi sebesar 93,45 dalam memprediksi pelanggan yang akan berlangganan berdasarkan kontak dari pelanggan yang menerima penawaran. Karim (2013) mengaplikasikan metode *Decision Tree* C4.5 dan *Naïve Bayes* untuk dibandingkan. Tujuannya untuk memprediksi apakah klien akan berlangganan deposito berjangka dengan dataset yang digunakan yaitu *bank direct marketing*. Hasil pengujian menunjukkan metode *decision tree* C4.5 lebih baik dibandingkan *naïve bayes* dengan akurasi yang didapat metode DT C4.5 sebesar 0,94 dan NB sebesar 0,87. Kemudian oleh Elsalamony (2014) melakukan penelitian tentang klasifikasi yang bertujuan untuk melihat kinerja metode-metode menggunakan teknik *data mining* yaitu teknik MLPNN, TAN, LR dan C5.0. Dari pengujian yang telah dilakukan, didapatkan teknik MLPNN memiliki performa yang baik dengan akurasi klasifikasi sebesar 90,92%. Menggunakan teknik TAN diperoleh akurasi klasifikasi sebesar 89,16%, kemudian dengan teknik LR diperoleh akurasi sebesar 90,09% dan teknik C5.0 sebesar 93,23%. Lalu penelitian yang dilakukan oleh Wisaeng (2013), melakukan perbandingan terhadap teknik klasifikasi metode *decission tree* dan *machine learning* dengan model yang digunakan yaitu JT48, LADT, RBFN, dan SVM. Hasil perbandingan menunjukkan model SVM unggul dengan akurasi sebesar 86,95. Disusul oleh model J48 sebesar 76,52, lalu model LADT sebesar 76,08, dan model RBFN sebesar 74,34. Penelitian yang dilakukan oleh Wisaeng ini menjadi referensi acuan untuk penelitian ini.

Dari uraian sebelumnya, penelitian ini mengusulkan melakukan perbandingan terhadap model dalam teknik klasifikasi pada *data mining*. Teknik klasifikasi yang digunakan yaitu *Support Vector Machine*, *AdaBoost*, *Naïve Bayes*, *Constant*, *KNN*, *Tree*, *Random Forest*, *Stochastic Gradient Descent*, dan *CN2 Rule*.

Model *Support Vector Machine* atau SVM merupakan *supervised learning* untuk masalah klasifikasi dan regresi (Shmilovici, 2009). SVM mampu untuk menyelesaikan masalah klasifikasi untuk data besar terutama pada permasalahan aplikasi multidomain di lingkungan *big data* (Suthaharan, 2016). Model AdaBoost merupakan model popular dalam *machine learning* yang implementasinya mudah dan dapat diterapkan dalam permalasalahan rekognisi dan klasifikasi. Namun untuk permasalahan klasifikasi, model ini mengkoreksi kesalahan yang dilakukan oleh pengklasifikasi lemah, sehingga rentan terhadap *overfitting* dibandingkan dengan model pembelajaran lain (Hu, dkk, 2008). Model *Naïve Bayes* merupakan model yang sederhana dan memiliki efisiensi yang cukup baik (Lewis, 1998). Model *naïve bayes* menwarkan klasifikasi kompetitif untuk kategorisasi teks dibandingkan model klasifikasi *data-driven* lainnya seperti jaringan saraf tiruan, SVM, dan KNN (Genkin, 2007). Model *Constant* merupakan model *classifier* yang melakukan prediksi terhadap distribusi kelas secara keseluruhan untuk setiap contoh secara sempurna (Flach, 2016). *K-Nearest Neighbor* atau KNN merupakan model klasifikasi yang dasar dan sederhana untuk distribus data. Klasifikasi KNN dikembangkan untuk melakukan analisis diskriminan ketika estimasi parametik reliabel dari kepadaan probabilitas sulit untuk ditentukan, kemudian US Air Force School of Aviation Medicine memperkenalkan metode non-parametrik untuk klasifikasi pola yang sejak saat itu dikenal menjadi aturan ketetanggaan (KNN) (Fix, & Hodges, 1951). Metode KNN telah berhasil melakukan proses klasifikasi pada permasalahan

pemilihan makanan sehat (Afandie, dkk, 2014) dan pemilihan bibit unggul sapi di Bali (Ekaristio, dkk, 2015).

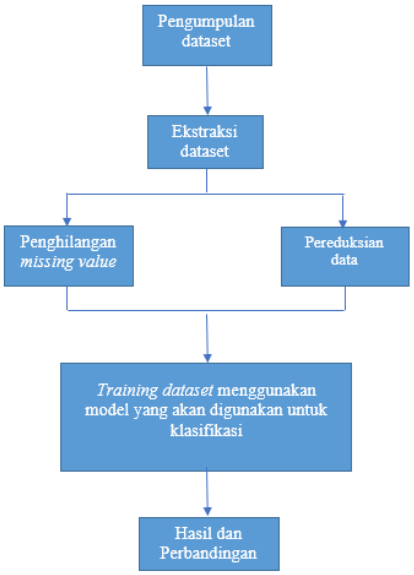
Model *Tree* dalam kaitannya dengan struktur data merupakan tipe data yang mensimulasikan struktur hierarkis *tree* dengan nilai akar dan anak-anak sub-*trees* dengan simpul induk diwakili sebagai serangkaian tautan *node* yang didefenisikan secara rekursif sebagai sarana penyajian analisis data kompleks (Klas, & Schrefl, 1995). Dalam pengaplikasiannya terhadap klasifikasi pada kasus ini, menggunakan pohon keputusan (*decision tree*) yaitu pernyataan control bersyarat yang menggunakan grafik mirip pohon atau model keputusan dan kemungkinan atas konsekuensinya dalam pembelajaran mesin (Quinlan, 1987). Model *Random Forest* merupakan model kombinasi dari *tree* yang menggunakan vektor acak yang diambil secara terpisah dari vektor input, dan setiap *tree* memberikan klas populernya untuk mengklasifikasikan vektor masukan (Breiman, 1999). Dengan kata lain, model ini menggunakan fitur yang dipilih secara acak atau kombinasi fitur disetiap simpul untuk membangkitkan *tree* (Pal, 2005). Model *Stochastic Gradient Descent* atau SGD merupakan model dalam *deep learning* yang mengoptimasi fungsi dengan mengikuti gradient yang memiliki *nosy* dengan ukuran langkah yang menurun (Mandt, dkk, 2017). Dan terakhir model *CN-2 Rule* merupakan model *association rule* yang menginduksi *rule* klasifikasi yang terurut sebagai pencarian heuristiknya (Clark, & Boswell, 1991). Kaitannya dengan klasifikasi, penambahan *rule* hanya ada satu target yang telah ditentukan.

Perbandingan teknik klasifikasi ini bertujuan untuk menentukan model dalam *data mining* yang memiliki akurasi terbaik untuk mengklasifikasi target yang sesuai pada *dataset Bank Direct Marketing*.

2. METODOLOGI PENELITIAN

Dalam penelitian ini mengusulkan perbandingan metode klasifikasi dalam *data mining* untuk *Bank Direct Marketing*. Adapun metodologi yang digunakan dalam penelitian ini digambarkan pada Gambar 1.

Pada Gambar 1, dijelaskan bahwa hal yang pertama dilakukan yaitu pengumpulan *dataset bank direct marketing* lalu mengekstraksi datanya. Kemudian dilakukan *preprocessing* dengan dua tahap yaitu proses penghilangan *unknown value* menggunakan *impute* pada *tool* Orange dan pemilihan fitur (*feature selection*) dengan menggunakan PCA. Selanjutnya dilakukan perbandingan metode klasifikasi dalam data mining. Setelah dilakukan perbandingan, kemudian dilakukan pegujian dengan data uji mengguunakan *10-fold cross validation*.



Gambar 1. Alur metodologi pada penelitian ini

2.1. *Dataset Bank Direct Marketing*

Sebelum dilakukan proses klasifikasi, dataset *Bank Direct Marketing* di ekstraksi terlebih dahulu yang diambil dari UCI *Repositoty* (Link : https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). Atribut awal pada dataset ini berjumlah 21 dengan 1 atribut tujuan dan memiliki 45.211 data instance. Deskripsi dari dataset dijelaskan pada Tabel 1.

Table 1. Atribut yang terdapat pada dataset *bank direct marketing* untuk mengklasifikasi algoritma

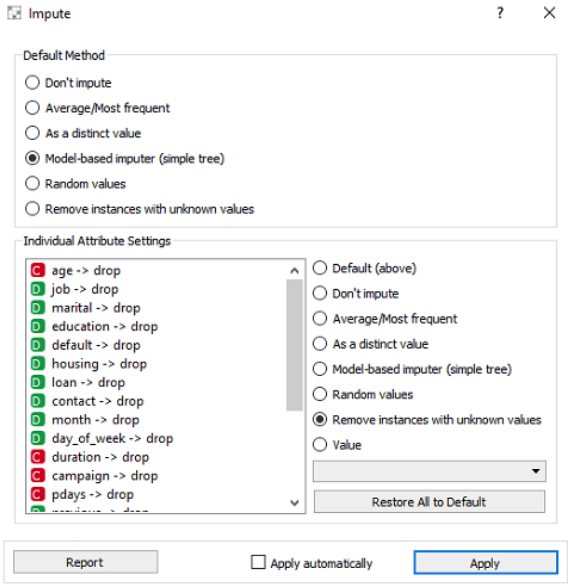
ID	Atribut	Type	Values	Descriptions
1	**Age**	**Numeric**	**Real**	**Age the contact date (≥18)**
2	**Job**	**Categorical**	**Admin, Unknown, Unemployed, Management, Housemaid, Entrepreneur, Student, Blue-collar, Self-employed, Retired, Technican, Services.**	
3	**Marital**	**Categorical**	**Maried, Diforced, Single, Widowed.**	
4	**Education**	**Categorical**	**Secondary, Unknown, Primary, Tertiary**	
5	**Default**	**Binary**	**Yes, No**	**Yes or No**
6	**Balance**	**Numeric**	**Real**	**In euro Currency**
7	**Housing**	**Binary**	**Yes, No**	**Yes or No**
8	**Loan**	**Binary**	**Yes, No**	**Yes or No**
9	**Contact**	**Categorical**	**Unknown, Telephone, Celullar**	
10	**Day**	**Numeric**	**Real**	**Referring to when the contact was made**

ID	Atribut	Type	Values	Descriptions
11	**Month**	**Categorical**	**Jan,Feb,., Nov, Dec,**	
12	**Duration**	**Numeric**	**Real**	**Of the contact (in seconds)**
13	**Campaign**	**Numeric**	**Real**	
14	**Pday**	**Numeric**	**Real**	
15	**Previous**	**Numeric**	**real**	
16	**Poutcome**	**Categorical**	**Unknown, Failure, Success**	

2.2. *Preprocessing*

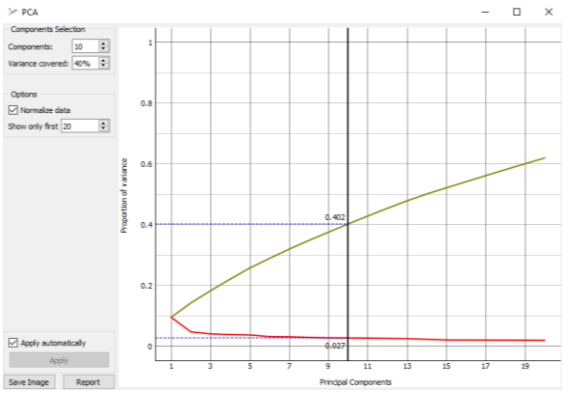
Pada *preprocessing*, dataset bank direct marketing ini memiliki *missing value* pada data. Adaya *missing value* pada data *instance* akan mengganggu berjalannya proses klasifikasi untuk *bank direct marketing*. Beberapa model dalam data mining untuk klasifikasi, tidak dapat berproses karena adanya *missing value.* Untuk itu diperlukan adanya penghilangan *missing value* pada data *instance*.

Missing value pada dataset bank direct marketing ini dihilangkan dengan menggunakan *tool impute* pada *software opensource* Orange[1]. *Impute* pada Orange berfungsi untuk mengganti *unknown value* pada dataset. Beberapa algoritme dan visualisasi Orange tidak dapat menangani nilai yang tidak dikenal dalam data. Pada Gambar 2, *widget impute* melakukan apa yang dinamakan *statisticians* imputasi yaitu menggantikan nilai yang hilang berdasarkan nilai yang dihitung dari data atau yang ditetapkan oleh pengguna. Imputasi defaultnya adalah 1-NN).



Gambar 2. *Widget impute*

Kemudian pada dataset ini juga dilakukan seleksi fitur dengan menggunakan PCA. *Principal Component Analysis* (PCA) dalam banyak cara membentuk dasar untuk analisis data dengan multivarian. Beberapa tujuan dari PCA yaitu menemukan hubungan antar objek. Dan tujuan dari PCA dari sisi lain yaitu mereduksi data. Pereduksian data digunakan pada saat sejumlah data besar dapat di dekati oleh struktur model yang cukup kompleks (Wold, 1987). Pada penelitian ini, PCA digunakan untuk mereduksi data.



Gambar 3. Reduksi menggunakan PCA

Dataset bank direct marketing yang memiliki 21 atribut dengan 1 atribut tujuan ini akan dilakukan seleksi fitur berdasarkan faktor yang mempengaruhi target dari tujuan. Pada Gambar 3, atribut yang berjumlah 20 pada *bank direct marketing* direduksi menjadi 10 atribut menggunakan *software open source* untuk data mining yang sebelumnya telah dilakukan pengujian *proporstion of variance*. Pereduksian atribut pada data bertujuan untuk mempercepat waktu komputasi model terhadap pengklasifikasian metode terhadap dataset *bank direct marketing*.

2.3. Model Klasifikasi

Dalam menganalisa performa dari beberapa teknik klasifikasi, maka dilakukan perbandingan metode dalam data mining baik metode *decission tree* maupun metode *machine learning* untuk memilih metode yang terbaik dengan akurasi yang tinggi dalam mengklasifikasi *dataset Bank Direct Marketing*. Menggunakan model klasifkasi yang ada pada *software opensource* dalam data mining yaitu berupa *constant, adaboost, stochastic gradient descent, k-nn, cn-2 rule inducer, svm, naive bayes, random forest* dan *tree.*

Constant pada *tool* data mining digunakan untuk memprediksi kelas yang paling sering muncul atau nilai rata-rata dari data yang telah dilatih. *Constant* menghasilkan model yang selalu memprediksi mayoritas untuk klasifikasi dan nilai rata-rata untuk regresi. Untuk klasifikasi, ketika ada dua atau lebih kelas mayoritas, *classifier* memilih kelas prediksi secara acak, tetapi selalu mengembalikan kelas yang sama untuk contoh tertentu.

AdaBoost (kependekan dari widget "*Adaptive boosting*") adalah algoritme *machine learning* yang berasal dari ide Yoav Freund dan Robert Schapire.

Model *AdaBoost* merupakan model klasifikasi *ensemble* dari meta algoritma yang menggabungkan *learning* yang '*weak*' untuk dilatih secara '*hardness*' dari setiap data sampel yang dilatih untuk meningkatkan kinerjanya.

Stochastic Gradient Descent menggunakan *gradien stochastic* yang meminimalkan fungsi kerugian yang dipilih dengan fungsi linear. Algoritma ini mendekati *gradien* yang benar dengan mempertimbangkan satu sampel pada suatu waktu, dan secara bersamaan memperbarui model berdasarkan *gradien* fungsi kerugian.

K-Nearest Neighbor menggunakan algoritma kNN yang mencari *k* dari contoh pelatihan terdekat dalam ruang fitur dan menggunakan rata-rata dari pencarian *k* sebagai prediksi.

Algoritma *CN-2 Rule* merupakan teknik klasifikasi yang dirancang untuk induksi yang efisien berdasarkan aturan yang mudah dipahami dan sederhana dari bentuk kondisional.

Support Vector Machine (SVM) merupakan teknik pada *machine learning* yang memisahkan ruang atribut dengan *hyperplane*, sehingga memaksimalkan *margin* antara *instance* dari suatu kelas dengtan nilai kelas. Pada *tool orange* berbasis Phyton, penerapan implementasi SVM cukup populer yang diambil dari paket LIBSVM. *Widget* dari SVM pada *tool* adalah antarmuka dari pengguna grafisnya.

Naïve Bayes merupakan probabilistik yang tergolong cepat dan sederhana berdasarkan pada teorema Bayes dengan asumsi fitur dapat berdiri sendiri. NB hanya dapat digunakan untuk mengklasifikasikan data.

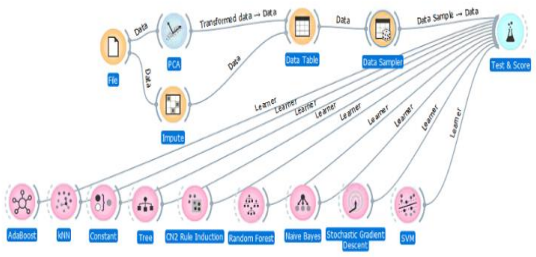
Random forest adalah metode *ensamble* dalam hal *learning* yang digunakan untuk klasifikasi, regresi, dan *task* lainnya. *Random forest* pertama kali diusulkan oleh Tin Kam Ho dan dikembangkan lebih lanjut oleh Leo Breiman dan Adele Cutler (Breman, 2001). Kinerja *random forest* diadaptasi dari *decission tree*, dengan setiap *tree* dikembangkan dari sampel *bootstrap* berdasarkan data latih. Ketika mengembangkan *tree*, subset atribut diambil secara acak dari atribut terbaik untuk dipilih secara *split*. Akhir model dari *random forest* didasarkan pada hasil dari keseluruhan subset *tree* yang telah dikembangkan.

Tree merupakan algoritma sederhana yang membagi data dari *node* ke *node* berdasarkan pembagian kelas. *Tree* lebih dahulu ditemukan daripada *random forest*. *Tree* pada *tool orange* dirancang secara *in-house* dan dapat menangani dataset diskrit maupun berkelanjutan.

Defenisi model klasifikasi yang digunakan dalam penelitian ini diambil dari dokumentasi referensi pada *software* Orange https://docs.orange.biolab.si/3/data-mining-library/#reference .

3. PENGUJIAN

Setelah dataset bank direct marketing diekstraksi, untuk membangun algoritma klasifikasi pada dataset diperlukan *software open source* data mining. *Software* yang digunakan yaitu *tool* dari Orange berdasarkan pemrograman Phyton dan dibawah lisensi GNU. Kemudian untuk tahap pengujian, dilakukan pengirisan data (data iris) dari data sebelumnya yang berjumlah 45.211 data *instance* menjadi 4.188 data *instance*. Hal ini dilakukan untuk mengurangi tingkat kompleksitas waktu pada saat dilakukan pengujian. Tahap selanjutnya setelah proses ekstraksi *dataset* yaitu melakukan penghilangan *missing value* menggunakan *tool impute* dan melakukan reduksi data dengan *tool* PCA pada *software* Orange. Setelah data tereduksi menjadi 10 variabel dan *missing value* pada data telah hilang, kemudian yang dilakukan yaitu melakukan pengujian dengan data uji menggunakan konsep *10-fold cross validation*. Data yang telah dilakukan *preprocessing*, dilakukan proses pengklasifikasian terhadap target dari dataset bank direct marketing menggunakan model yang ada pada *software* Orange. Pada Gambar 4, menjelaskan proses yang dilakukan mulai dari penghilangan *missing value* sampai pada proses pengklasifikasian dengan model pada *tool*.



Gambar 4. Proses klasifikasi *bank direct marketing*

4. HASIL DAN PEMBAHASAN

Pada tahap ini merepresentasikan hasil yang diperoleh setelah dilakukan pengujian terhadap sembilan model klasifikasi dalam data mining. Hasil klasifikasi model dalam *data mining* terhadap target untuk *bank direct marketing* ditunjukkan pada Tabel 2.

Tabel 2. Hasil eksperimen terhadap model klasifikasi untuk *bank direct marketing*

Methods	CA	Precission	Recall
AdaBoost	0.945	0.948	0.945
CN-2 Rule	0.955	0.947	0.955
Random Forest	0.971	0.945	0.971
Tree	0.972	0.945	0.972
KNN	0.971	0.945	0.971
Naïve Bayes	0.972	0.945	0.972
SVM	0.671	0.948	0.671

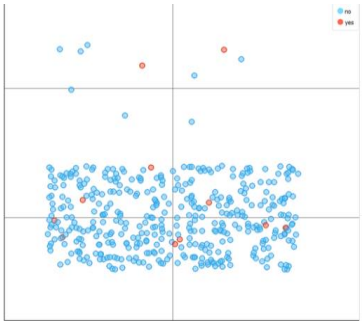
Methods	CA	Precission	Recall
Constant	0.972	0.945	0.972
SGD	0.972	0.945	0.972

Pada Tabel 2, hasil pengujian terhadap model klasifikasi ditunjukkan dengan menggunakan metode *scoring*. Metode *scoring* yang digunakan untuk menghitung akurasi klasifikasi ini berasal dari *tool* Orange sebagai *software* pengujiannya. *CA*, *Precission*, dan *Recall* merupakan scoring method yang digunakan untuk pengujian ini. *CA* digunakan untuk mengitung akurasi subset. *Precission* digunakan untuk akurasi klasifikasi secara intuitif, dengan nilai terbaik adalah 1 dan terburuk adalah 0. *Recall* digunakan untuk mengukur rasio pengklasifikasian.

Hasil *scoring* menunjukkan nilai akurasi terbaik yaitu metode *Tree* dengan nilai *CA* sebesar 0,972, *Precission* sebesar 0,945, dan *Recall* sebesar 0,972. Nilai akurasi yang sama juga didapatkan oleh metode *Naive Bayes*, *Support Vector Machine*, *Constant*, dan *Stochastic Gradient Descent*.

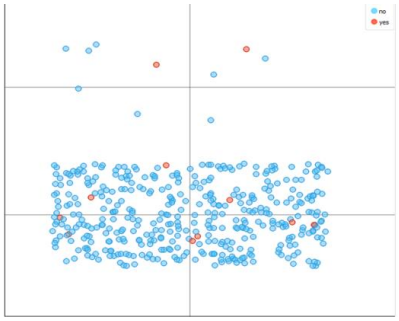
Kemudian hasil *scoring* dengan metode lain menunjukkan bahwa metode *Random Forest* dan metode *K-NN* memiliki nilai akurasi *CA* sebesar 0,971, *Precission* sebesar 0,945, dan *Recall* sebesar 0,971. Dilanjutkan dengan metode *CN-2 Rule Inducer* dengan nilai akurasi *CA* sebesar 0,955, *Precission* sebesar 0,947, dan *Recall* sebesar 0,955. Lalu metode *AdaBoost* memiliki akurasi dengan nilai *CA* sebesar 0,945, *Precission* sebesar 0,948, dan *Recall* sebesar 0,945. Metode SVM pada penelitian ini memiliki akurasi nilai *CA* sebesar 0,671, dengan nilai *Precission* sebesar 0,948 dan *Recall* sebesar 0,671. Metode SVM pada pengujian ini rendah dikarenakan metode ini tidak memiliki kemampuan *adaptive* ketika melakukan perubahan pada *preprocessing* maupun pengujian saat menggunakan metode *10-fold cross validation*.

Setelah melakukan *scoring method* untuk menentukan hasilnya setelah dilakukan pengujian, hasil pengujian yang telah dilakukan dapat dilihat dari visualisasi yang dihasilkan atas kinerja model yang digunakan.



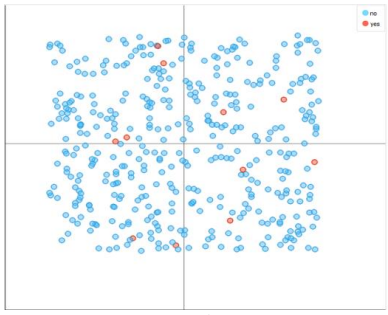
Gambar 5. Visualisasi hasil klasifikasi dengan model *AdaBoost*

Hasil visualisasi yang tampak pada Gambar 5, menunjukkan model *AdaBoost* mampu untuk mengklasifikasi target tujuan untuk *bank direct Marketing*. Namun model *AdaBoost* ini tidak mampu menghasilkan bentuk visualisasi yang bagus dalam mengklasifikasi target tujuan.



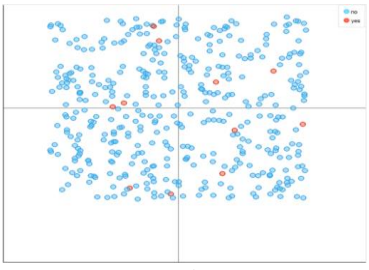
Gambar 6. Visualisasi hasil klasifikasi dengan model *CN-2 Rule*

Hasil visualisasi pada Gambar 6, memperlihatkan bahwa model *CN-2 Rule* memiliki kemiripan visualisasi dengan model *AdaBoost*. Sehingga bentuk visualisasi dari model ini juga tidak dapat menghasilkan bentuk visualisasi yang terbaik.



Gambar 7. Visualisasi hasil klasifikasi dengan model *Random Forest*

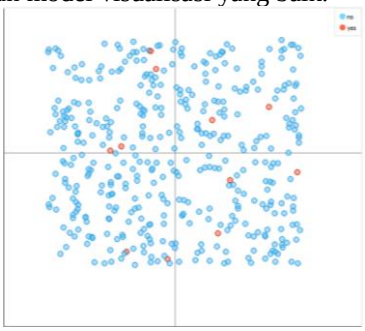
Bentuk visualisasi pada Gambar 7 yang dibentuk oleh model *Random Forest* terlihat menyatu antar target tujuan namun mampu untuk menghasilkan nilai klasifikasi. *Random Forest* juga tidak dapat menghasilkan bentuk visualisasi yang terbaik dalam mengklasifikasi target tujuan untuk *bank direct marketing*.



Gambar 8. Visualisasi hasil klasifikasi dengan model *Tree*

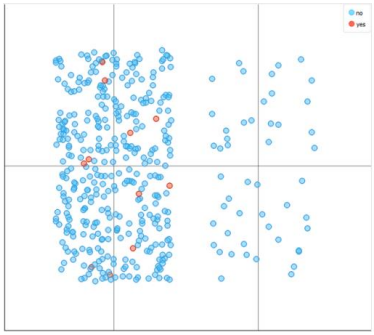
Pada Gambar 8, menunjukkan hasil visualisasi model *Tree* terhadap hasil klasifikasi untuk *bank direct marketing*. Meskipun akurasi nilai

dari model *Tree* termasuk tinggi, namun pada pengimplementasian model ini tidak mampu untuk menghasilkan model visualisasi yang baik.



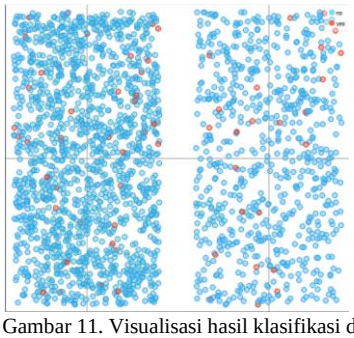
Gambar 9. Visualisasi hasil klasifikasi dengan model *K-Nearest Neighbor*

Visualisasi model *K-Nearest Neighbor* (KNN) pada Gambar 9, menunjukkan kemiripan dengan model *Random Forest* dikarenakan hasil akurasi yang juga sama. Oleh karena itu, model KNN ini juga tidak mampu menghasilkan visualisasi klasifikasi yang baik untuk memperlihatkan target tujuan klasifikasi pada *bank direct marketing*.



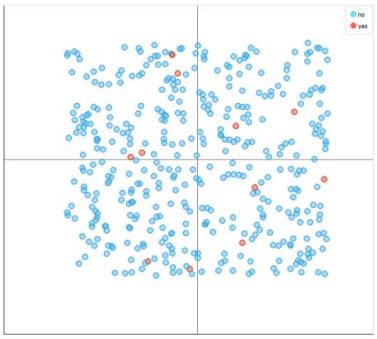
Gambar 10. Visualisasi hasil klasifikasi dengan model *Naïve Bayes*

Hasil klasifikasi model *Naïve Bayes* berdasarkan menunjukkan kemampuan model ini untuk mengklasifikasi target tujuan berdasarkan nilai akurasi. Namun pada Gambar 10, terlihat bahwa model ini tidak mampu memberikan visualisasi yang baik dalam mengklasifikasi target tujuan dengan melihat salah satu target tujuan yaitu *yes* hanya berada disatu kolom saja.



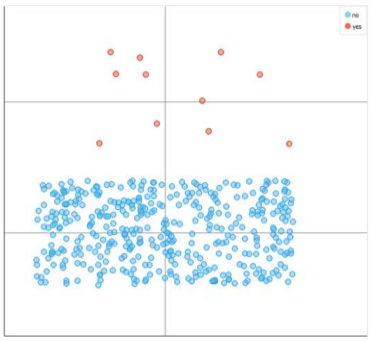
Gambar 11. Visualisasi hasil klasifikasi dengan model *Support Vector Machine*

Pada Gambar 11, model *Support Vector Machine* (SVM) menunjukkan hasil visualisasinya. Terlihat bahwa SVM memiliki kelemahan dalam mengklasifikasi target tujuan pada *bank direct marketing* sesuai dengan nilai akurasinya yang rendah. Hasil visualisasi model ini terlihat lebih ramai sehingga sulit untuk melihat dengan jelas target tujuan yang dicapai.



Gambar 12. Visualisasi hasil klasifikasi dengan model *Constant*

Berdasarkan Gambar 12, model *Constant* mampu untuk melakukan klasifikasi target tujuan untuk *bank direct marketing*. Hasil visualisasi yang terlihat pada Gambar 12, juga memiliki kemiripan bentuk visualisasi dengan model *Tree* dan model KNN.



Gambar 13. Visualisasi hasil klasifikasi dengan model S*tochastic Gradient Descent*

Hasil visualisasi pada Gambar 13, untuk model *Stochastic Gradient Descent* (SGD) menghasilkan bentuk visualisasi yang jelas. Meskipun nilai akurasi model ini sama dengan model *Tree*, *Naïve Bayes*, dan *Constant* namun model SGD menghasilkan visualisasi yang terbaik dan jelas berdasarkan bentuk klasifikasinya. Dari gambar 13 terlihat, kemampuan SGD dalam membelah hasil klasifikasi sesuai target tujuan untuk *bank direct marketing*.

Dari pengujian yang telah dilakukan pada tabel 2 dan penjelasan masing-masing model, maka pada tabel 3 dijelaskan tentang kelebihan dan kelemahan model-model yang diusulkan pada penelitian ini. Adapun kelemahan dan kelebihan kesembilan model yang diusulkan pada penelitian in dijelaskan pada Tabel 3 sebagai berikut:

Tabel 3. Perbandingan model klasifikasi untuk *bank direct marketing*

	Metode	Kelebihan	Kelemahan
1	AdaBoost	a. Mudah dalam implementasi b. Dapat diimplementasikan dalam kasus klasifikasi dan pengenalan c. Memiliki nilai akurasi yang tinggi dalam mengklasifikasi target tujuan	a. Dalam kasus ini, Adaboost memiliki visualisasi yang kurang bagus untuk mengklasifikasi target tujuan b. Dalam kasus klasifikasi, model ini dinilai rentan terhadap overfitting
2	CN-2 Rule	Pada model ini harus melakukan evaluasi *rule* yang telah ditemukan untuk dapat memutuskan mana yang terbaik. Aturan kualitas yang memungkinkan adalah akurasinya dalam data yang telah dilakukan pelatihan. Pada pengujian ini, model CN2-Rule memiliki nilai akurasi yang cukup baik.	Masalah pada kemungkinan menemukan akurasi yang baik, model ini cenderung memilih rule yang sangat spesifik. Karena kemungkinan menemukan aturan dengan akurasi tinggi pada data latih akan meningkatkan *rule* untuk lebih spesifik lagi.
3	Random Forest	a. Untuk banyak dataset, model pembelajaran ini menghasilkan *classifier* yang sangat akurat. b. Untuk dataset yang berukuran besar, model ini berjalan dengan efisien c. Model ini efektif ketika berhadapan dengan data yang memiliki *missing value* dengan tetap mempertahankan akurasi bahkan ketika sebagian besar data hilang	Pada jenis data yang termasuk variable kategori dengan jumlah tingkat yang berbeda, model Random Forest memiliki nilai bias untuk banyak level. Untuk itu, nilai variable Random Forest tidak dapat diandalkan untuk jenis data yang memiliki variable kategori yang berbeda.
4	Tree	a. Merupakan model *over-simple* b. Digunakan untuk jenis data dengan nilai inputan berupa nilai diskrit c. Pemilihan fitur dilakukan secara otomatis, sehingga lebih cepat dalam pemprosesan dan konfigurasi parameter lebih sedikit d. Pada pengujian ini, model Tree memiliki nilai akurasi yang tinggi	a. Model *tree* mengantisipasi jalan buntu, namun ketika salah dalam menentukan langkah, maka akan mengakibatkan kesalahan yang mengganggu proses yang telah ada. b. Kecenderungan model ini, hanya melihat data historis dari jalur yang pernah dilalui, sehingga melemahkan situasi perubahan.
5	K-NN	a. Model ini tahan terhadap data yang memiliki noisy b. Efektif terhadap data latih yang besar	a. Model ini perlu menentukan nilai parameter K b. Biaya komputasi cukup tinggi karena perlu menghitung dulu jarak setiap contoh pencarian untuk semua data latih.
6	Naïve Bayes	a. Sangat simple, mudah untuk digunakan dan cepat b. Membutuhkan lebih sedikit data pelatihan c. Menangani data yang kontinyu maupun diskrit d. Model ini juga dapat digunakan untuk prediksi probabilistik	a. NB membuat asumsi yang sangat kuat yang berdampak pada bentuk distribusi data b. Ketika data memiliki *missing value*, perlu dilakukan perkiraan nilai kemungkinan dengan pendekatan frequentist untuk dapat menghasilkan probabilitas.
7	SVM	a. Menggunakan parameter regulasi, untuk menghindari terjadinya *over-fitting* b. Menggunakan metode Kernel c. Tidak memiliki lokal minimum	a. Pada pengujian ini, nilai akurasi yang didapatkan model ini cenderung rendah. Ini disebabkan pada dataset ini terdapat missing value, dan model ini rendah dalam menangani data yang memiliki missing value. b. Teori kernel yang diadopsi oleh model ini hanya mencakup parameter untuk nilai tertentu dari regulasi parameter dan pilihan kernel. Sedang Kernel
8	Constant	a. Cocok digunakan untuk model prediktif b. Untuk kasus klasifikasi, model ini melakukan validasi dan evaluasi terhadap model data berdasarkan spesifikasi dan karakteristik data	Model ini membutuhkan tes diskriminasi untuk mengetahui apakah hasil akurasi, namun terkadang tidak dapat memastikan apakah prediksi tersebut akurat dan tanpa bias
9	SGD	a. Mampu meningkatkan kinerja generalisasi masalah berskala besar b. Mampu diterapkan pada data dengan jumlah besar dan memiliki missing value c. Waktu komputasi relatif lebih singkat meskipun datanya besar	a. Pada data dengan skala kecil, ketersediaan data menjadi batasan, bukan waktu komputasi. b. SGD memiliki sensitivitas terhadap penskalaan fitur c. Model ini membutuhkan beberapa parameter seperti parameter regulasi dan jumlah iterasi

5. KESIMPULAN

Banyak algoritma yang diusulkan untuk klasifikasi pada dataset *bank direct marketing*. Namun pada penelitian ini mengusulkan model yang ada pada *tool software open source* dalam *data mining*. Setelah dilakukan pengujian dengan menggunakan *tool*, didapat hasil klasifikasi dengan akurasi yang tinggi pada model *Tree, NB, Constant* dan *SGD.* Disusul dengan model *random forest, KNN, CN-2 Rule Inducer, AdaBoost,* dan SVM.

Kesembilan model yang diusulkan memiliki karakteristik masing-masing model berdasarkan spesifikasi datanya. Berdasarkan pada permasalahan ini, SVM yang cenderung sering diimplementasikan untuk kasus klasifikasi pada beberapa penelitian, yang pada permasalahan ini mendapatkan akurasi klasifikasi yang rendah yaitu sebesar 0,671. Ini disebabkan karena pada dataset BDM ini memiliki *missing value* dimana menjadi kekurangan model ini. Lalu model *AdaBoost* yang merupakan model dalam *machine learning* yang mudah untuk diimplementasikan ini memiliki akurasi yang cukup tinggi yaitu sebesar 0,945. Namun, untuk permasalahan ini model *adaboost* menghasilkan visualisasi yang belum bagus. Lalu model *CN-2 Rule*, memiliki nilai akurasi yang cukup tinggi dibandingkan model *adaboost* yaitu 0,955. Kelebihan model ini yaitu melakukan evaluasi *rule* untuk mendapatkan solusi terbaik, namun kemungkinan untuk menemukan solusi terbaik akan membuat *rule* akan lebih spesifik lagi. Kemudian model *K-Nearest Neighbor* dan *Random Forest* yang memiliki nilai akurasi sama yaitu 0,971. Keduanya cocok digunakan untuk dataset dengan jumlah data besar sesuai dengan kelebihan kedua model ini, namun kedua model ini memiliki kelemahan dari segi waktu komputasi dan penentuan nilai variabel ketika pada data latih terdapat variabel kategori yang berbeda. Dan selanjutnya yaitu keempat model yang menghasilkan nilai akrasi klasifikasi yang tinggi, yaitu model *Tree*, *Naïve Bayes*, *Constant*, dan *Stochastic Gradient Descent* sebesar 0,972. Model *tree* dan NB yang merupakan model dengan implementasi yang mudah dan cepat, namun model tree cenderung terhadap data historis dan model NB cenderung membuat asumsi yang sangat kuat yang akan mempengaruhi distribusi data. Model *constant* lebih cocok digunakan untuk kasus prediksi, namun juga bisa untuk klasifikasi. Untuk kasus klasifikasi, model ini melakukan evaluasi terhadap model data dan spesifikasinya, tetapi model ini dalam proses menentukan solusi terbaik membutuhkan tes diskriminasi. Dan terakhir yaitu SGD. Model ini merupakan model yang dirancang untuk mengoptimasi fungsi pencarian solusi yang dapat digunakan untuk data dengan skala jumlah besar. Namun model ini memiliki sensitivitas terhadap penskalaan fitur dan waktu komputasi yang sedikit lama.

Kemudian berdasarkan visualisasi yang dihasilkan dari pengujian yang sesuai dengan Tabel 2 menggunakan *tool* Orange, terlihat bahwa *Stochastic Gradient Descent* (SGD) memiliki visualisasi yang terbaik untuk mencapai target tujuan klasifikasi untuk *bank direct marketing* yaitu diterima (*yes*) dan ditolak (*no*). Dalam visualisasinya, SGD dapat membedakan hasil klasifikasi target tujuan secara jelas dibandingkan dengan model lainnya. Sementara SVM merupakan keterbalikannya. SVM dalam mengklasifikasi target untuk *bank direct marketing* ini tidak mampu dikarenakan hasil akurasi klasifikasi yang rendah dibandingkan metode lain sehingga visualisasi yang dihasilkan juga kurang baik. Untuk itu, dalam kasus klasifikasi target tujuan pada dataset *bank direct marketing* ini SGD unggul dalam nilai akurasi yang didapat sebesar 0,972 yang diikuti dengan visualisasi model setelah data dilatih.

DAFTAR PUSTAKA

AFANDIE, M. N., CHOLISSODIN, I., & SUPIANTO, A. A, 2014, Implementasi metode k-nearest neighbor untuk pendukung keputusan pemilihan menu makanan sehat. Repositori Jurnal Mahasiswa PTIIK UB, *3*(1), 1.

ANGGODO, Y.P., CAHYANINGRUM, W., FAUZIYAH, A. N., KHOIRIYAH, I.L., KARTIKASARI, O., CHOLISSODIN, I., 2017, Hybrid K-Means dan Particle Swarm Optimization untuk clustering nasabah kredit, Jurnal Teknologi Informasi dan Ilmu Komputer, hlm. 104-110.

BARTIK, V, 2009, Assosiation based classification for relational data and its use in web mining, IEEE Symposium on Computatioanl Intelligence and Data mining, pp. 252-258.

BREIMAN, L., 1999, Random forests – random features, Technical Report 567, Statistics Departement, University of California, Berkeley.

CLARK, P., & BOSWELL, R., 1991, Rule induction with CN2: Some recent improvements, In: Kodratoff Y.(eds) Machine Learning – EWSL-91, EWSL 1991, Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 482, Springer, Berlin, Heidelberg.

EKARISTIO, I., SOEBROTO, A. A., & SUPIANTO, A. A, 2015, Pengembangan sistem pendukung keputusan pemilihan bibit unggul sapi bali menggunakan metode k-nearest neighbor. Journal of Environmental Engineering and Sustainable Technology, *02*(01), 49–57.

ELSALAMONY, H. A., & ELSAYAD, A. M., 2013, Bank direct marketing based on neural

network, International Journal of Engineering and Advanced Technology, vol.2, pp. 392-400.

ELSALAMONY, H.A., 2014, Bank direct marketing analysis of data mining techniques, International Journal of Computer Applications, vol. 85, no.7.

GRZONKA, D., SUCHACKA, G., BOROWIK, B., 2016, Application of selected supervised classification methods to bank marketing campaign, Information Syatems in Management, vol.5 (1), pp. 36-48.

FIX, E., & HODGES, J. L., 1951, Discriminatory analysis, nonparametric discrimination: Consistency properties, Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas.

FLACH, P. A., Classifier calibration, In: C. Sammut, G.I., Webb (eds), Encyclopedia of machine learning and data mining, Springer, Boston, MA.

GENKIN, A., LEWIS, D. D., & MADIGAN, D., 2007, Large-scale Bayesian logistic regression for text categorization, Technometrics, vol.49, pp. 291-304.

HAO, Z., WANG, Z., & ZHANG, Y., 2009, Improved classification based on predictive associative rules, IEEE International Conference on System, Man and Cybernatics, pp. 1165-1170.

HU, W., HU, W., & MAYBANK, S., 2008, AdaBoost-based algorithm for network instrusion detection, IEEE Transactions On Systems, Man, and Cybernetics - Part B: Cybernatics, vol.38, no.2.

KARIM, M., & RAHMAN, R. M., 2013, Decission tree and naïve bayes algorithm for classification and generation of actionable knowledge for direct marketing, Journal of Software Engineering and Application, vol.6,pp.196-206.

KLAS, W., & SCHRELF, M., 1995, Metaclasses and their application: Data model tailoring and database integration. Springer.

LEWIS, D. D., 1998, Naïve (Bayes) at forty: The independence assumption in information retrieval, In European Conference on Machine Learning, pp. 4-15.

MANDT, S., HOFFMAN, M. D., & BLEI, D. M., 2017, Stochastic gradient descent as approximate Bayesian inference, Journal of Machine Learning Research, 18, 1-35.

NIU, Q., XIA, X., & ZHANG, L., 2009, Assosiation classification based on compactness of rules, International Workshop On Knowledge Discovery And Data Mining, pp. 245-247.

PAL, M., 2005, Random forest classifier for remote sensing classification, Internatonal Journal of Remote Sensing, 26:1, 217-222.

QUINLAN, J.R., 1987, Simplifying decision trees, International Journal of Man-Machine Studies 27, pp: 221-234.

SHMILOVICI, A., 2009, Support vector machine, In:Maimon O., Rokach L. (eds) Data mining and knowledge discovery handbook, Springer, Boston, MA.

SUTHAHARAN, S., 2016, Support vector machine, In: Machine leaning models and algorithms for big data classification, Integrated Series In Information Systems, vol.36, Springer, Boston, MA.

VAIDEHI, R., 2016, Predictive modelling to improve successs raate of bank direct marketing campaign, International Journal of Management and Bussiness Study, vol 6, pp. 22-24.

VIJAYAKUMAR, V., & NEDUNCHEZHIAN, R., 2012, A study on video data mining, International Journal of Multimedia Information Retrieval, vol 1, issue 3, pp 153-172.