2 (2019): 910-915
, Prosiding Seminar Nasional Matematika
ournal.unnes.ac.id/sju/index.php/prisma/
3-9189

Perbandingan Metode Klasifikasi Supervised Learning Pada Data Bank Customers Menggunakan Python

Fajar Sodik Pamungkas[a], Bayu Dwi Prasetya[b], Iqbal Kharisudin[a,b]
[a,b] *Universitas Negeri Semarang*

Email : fajarswodik@gmail.com

Abstrak

Makalah ini bertujuan untuk menganalisis dan membandingkan metode–metode pendekatan supervised learning dengan menggunakan studi kasus data churn modelling dari kaggle. Penelitian ini menggunakan program jupyter notebook berbahasa python. Langkah yang dilakukan adalah persiapan yaitu untuk menyiapkan modul yang digunakan untuk analisis supervised learning. kemudian pre-processing data yaitu melabeli data yang memiliki tipe data kategorik, setelah itu membagi data untuk data training dan tesing dengan perbandingan 9:1. Lalu dibuat model dan confussion matriknya untuk setiap metode. Metode yang digunakan adalah regresi logistik, K-nearest neighboor, naive bayes, super vector machine, dan random forest.

Berdasarkan hasil perhitungan akurasi metode supervised learning diperoleh nilai: 0,82 untuk metode regresi logistik, 0,839 untuk metode K-nearest Neighboor, 0,8 untuk metode super vector machine, 0,836 untuk metode naive bayes, 0,791 untuk metode decision tree, 0,862 untuk metode random forest. Jadi dilihat dari akurasinya maka metode random forest adalah metode terbaik untuk menganalisis data bank-customer dengan nilai akurasi 0,862.

Kata kunci: *supervised learning, regresi logistik, K-nearest neighboor, super vector machine, naive bayes, decision tree, random forest*

1. Pendahuluan

1.1. Latar Belakang

Data Mining merupakan kumpulan dari kegiatan yang meliputi pengumpulan dan pemakaian data masa lalu untuk menemukan pola atau hubungan dalam data yang berukuran besar. Output pada data mining tersebut dapat dijadikan pengambilan keputusan dimasa depan. Metode ini merupakan gabungan dari 4 disiplin ilmu yaitu visualisasi, statistik, basis data dan machine learning. Adapun klasifikasi merupakan salah satu ilmu yang terdapat pada machine learning. Klasifikasi merupakan salah satu metode yang dapat menangani big data. Terdapat dua metode dalam klasifikasi yaitu supervised learning dan unsupervised learning. Supervised Learning merupakan algoritma yang membangkitkan suatu fungsi yang memetakan input ke output yang diinginkan. Terdapat banyak metode yang ada dalam klasifikasi supervised learning diantaranya adalah diantaranya Regresi Logistik, *K-nearest Neighboor*, *Super Vector Machine*, *Naive Bayes*, *Decision Tree* dan *Random Forest*.

Kaggle merupakan situs web yang menyediakan banyak *dataset* yang digunakan untuk keperluan *data science*, salah satunya adalah data *bank customers*. Data *bank customers* ini memiliki 10000 data dengan 13 variabel *input* yaitu *rownames,CustID, SurName, CreditScore, Geography, Gender, Age, Tenure, Balance, NumProduct, HasCrCard, IsActiveMember, EstimatedSalary,* dengan 1 variabel target yaitu *Exited* yang berarti pelanggan akan menutup akun atau tidak.

Untuk menganalisis data apalagi data yang cukup besar diperlukan suatu alat untuk menghitung dan menganalisis supaya lebih efektif dan efisien. Python merupakan salah satu alat yang direkomendasikan untuk hal tersebut. Dalam beberapa tahun terakhir, Dukungan perpustakaan Python yang ditingkatkan (terutama panda) telah membuatnya menjadi alternatif yang kuat untuk tugas analisis data. Dikombinasikan

To cite this article:
Pertama, P., Kedua, P., & Ketiga, P. (2019). Klik di sini untuk menulis judul anda. *PRISMA, Prosiding Seminar Nasional Matematika 2*, 910-915

dengan kekuatan Python dalam pemrograman tujuan umum, itu adalah pilihan yang sangat baik sebagai bahasa tunggal untuk membangun aplikasi data-sentris (McKinney Wes, 2012).

1.2. Rumusan Masalah

Masing-masing metode klasifikasi *supervised learning* punya kelebihan, kekurangan, dan hasil ketepatanya masing-masing. Melalui data bank customers dari kaggle, bagaimana metode klasifikasi *supervised learning* bekerja dan bagaimana hasil ketepatanya

1.3. Tujuan

Dalam makalah ini akan dibahas mengenai ketepatan klasifikasi dari masing-masing analisis, dimana data yang digunakan adalah data bank customers yang diperoleh dari kaggle. Dari hasil ketepatan klasifikasi tersebut kemudian dibandingkan dan ditentukan mana metode klasifikasi tertepat.

2. Metode

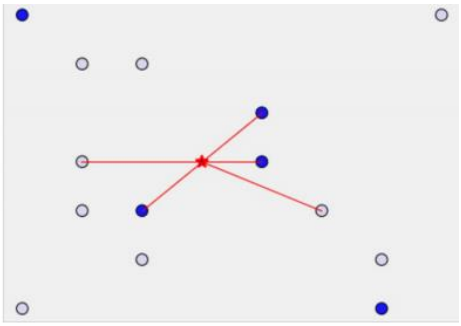
2.1. Regresi Logistik

Regresi Logistik adalah suatu analisis regresi yang digunakan untuk menggambarkan hubungan antara variabel respon dengan sekumpulan variabel prediktor dimana variabel respon bersifat biner atau dikotomus. Regresi logistik Biner digunakan saat variabel dependen merupakan variabel dikotomus (kategorik dengan 2 macam kagegori),. Regresi Logistik tidak memodelkan secara langsung variabel dependen (Y) dengan variabel independen (X), melainkan melalui transformasi variabel dependen ke variabel logit yang merupakan natural log dari odds rasio (Fractal, 2003). metode ini cukup tahan (*robust*) untuk dapat diterapkan dalam berbagai skala/keadaan data (Tatham et. al, 1998). Model regresi logistik multivariate dengan k variabel prediktor adalah :

$$\pi(x) = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right.}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right.}$$

2.2. K-Nearest Neighboor

Metode KNN merupakan salah satu metode klasifikasi yang mudah. Metode ini bekerja dengan mencari k pola (diantara semua pola latih disemua kelas) yang terdekat dengan pola masukan kemudian menentukan kelas keputusan berdasarkan jumlah pola terbanyak (Suyanto, 2018). Proses pelatihan KNN menghasilkan k yang memberikan akurasi tertinggi dalam menggeneralisasi data yang akan datang. Masalahnya, sampai saat ini k tidak dapat ditentukan secara matematik. Jadi proses pelatihan proses pelatihan pada dasarnya adalah melakukan observasi terhadap sejumlah k sampai dihasilkan k yang paling optimum.

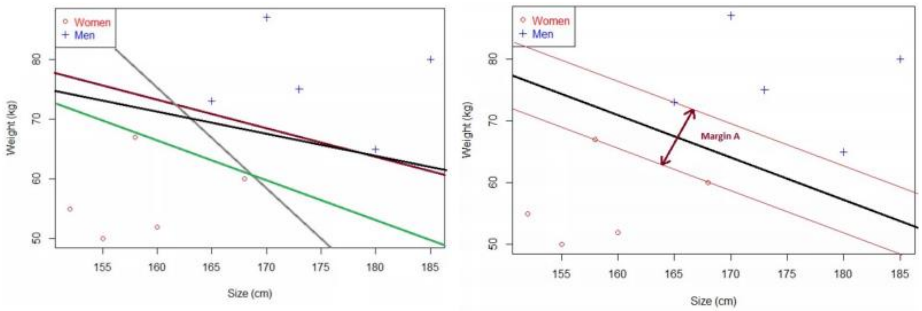


Gambar 1. Visualisasi *K-Nearest Neighboor*

2.3. Super Vector Machine

Metode SVM SVM bertujuan untuk mencari hyperplane yang optimal. Hyperplane yang dapat membagi kedua class dengan jarak margin terjauh antar class. Margin adalah jarak antara hyperplane tersebut dengan pola terdekat dari masing-masing class. Instance yang paling dekat ini disebut sebagai support vector. Pada garis merah yang berada di atas garis hitam tebal dapat instance dengan tanda "+" yang menjadi support vector untuk kelas Men. Sedangkan pada garis merah di bawah garis hitam tebal terdapat instace dengan tanda "o" yang menjadi support vector untuk kelas Women. Sehingga dapat

disimpulkan bahwa tujuan utama SVM adalah mencari hyperplane terbaik dengan bantuan support vector dari masing-masing class sehingga akhirnya didapat hyperplane optimal.



Gambar 2. (a) Visualisasi hyperlane tidak maksimal (b) Visualisasi hyperlane maksimal

2.4. Naive bayes

Metode Naive Bayes adalah metode yang menggunakan sebuah teorema kuno warisan abad ke 18 yang ditemukan oleh Thomas Bayes (Suyanto, 2018). Dalam teorema tersebut suatu probabilitas bersayarat dinyatakan sebagai berikut:

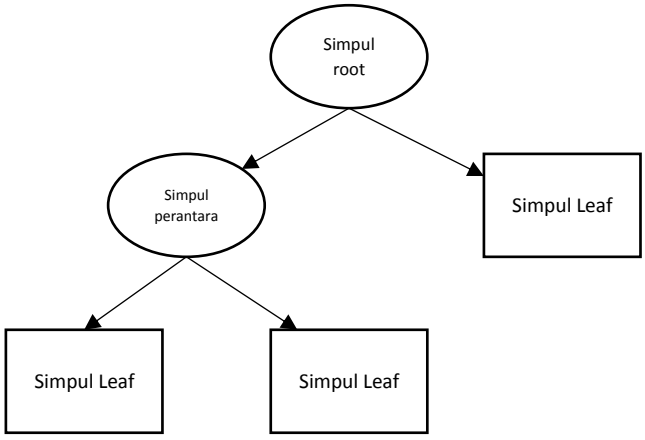
$$P(X|H) = \frac{P(X|H)P(H)}{P(X)}$$

Dimana X adalah bukti, H adalah hipotesis, $P(H|X)$ adalah probabilitas bahwa hipotesis H benar untuk bukti X atau dengan kata lain $P(H|X)$ merupakan probabilitas posterior H dengan syarat X. Dalam bidang machine learning X adalah sebuah objek data, H adalah hipotesis bahwa X adalah kelas C. Secara spesifik, dalam masalah klasifikasi dapat dihitung $P(H|X)$ sebagai probabilitas bahwa hipotesis benar untuk *tuple* X, dengan kata lain $P(H|X)$ adalah probabilitas bahwa *tuple* X berada dalam kelas C.

2.5. *Decision Tree*

Decision tree yang dikenal juga sebagai *top-down induction of decision trees* (TIDIDT) adalah teknik supervised learning yang membangun representasi aturan klasifikasi berstruktur sekuensial hirarki dengan cara mempartisi himpunan data latih secara rekursif (Suyanto, 2018). Teknik ini menghasilkan pohon keputusan yang berupa n-ary *branching tree* yang mereprensentasikan suatu aturan klasifikasi. Beberapa teknik decision tree adalah: *classification and regression tree* (CART), *itterative dychotomizer* (ID3), *C4.4 Quinlan, C5.0 Quinlan, Cubist Quinlan, Assistant Cestnik.* digunakan untuk pengenalan pola dan termasuk dalam pengenalan pola secara statistik.decision tree dibentuk dari 3 tipe simpul

- Simpul leaf memuat suatu keputusan akhir atau kelas target untuk suatu decision tree
- Simpul root adalah titik awal dari decision tree
- Setiap simpul perantara berhubungan dengan suatu pertanyaan atau pengujian



Gambar 3. Visualisasi Decision Tree

2.6. Random Forest

Metode random forest adalah pengembangan dari metode CART, yaitu dengan menerapkan metode bootstrap aggregating (bagging) dan random feature selection (Breiman 2001). Dalam random forest, banyak pohon ditumbuhkan sehingga terbentuk hutan (forest), kemudian analisis dilakukan pada kumpulan pohon tersebut. Pada gugus data yang terdiri atas n amatan dan p peubah penjelas, random forest dilakukan dengan cara (Breiman 2001; Breiman & Cutler 2003):

- Lakukan penarikan contoh acak berukuran n dengan pemulihan pada gugus data. Tahapan ini merupakan tahapan bootstrap.
- Dengan menggunakan contoh bootstrap, pohon dibangun sampai mencapai ukuran maksimum (tanpa pemangkasan). Pada setiap simpul, pemilihan pemilah dilakukan dengan memilih m peubah penjelas secara acak, dimana m << p. Pemilah terbaik dipilih dari m peubah penjelas tersebut. Tahapan ini adalah tahapan random feature selection.
- Ulangi langkah 1 dan 2 sebanyak k kali, sehingga terbentuk sebuah hutan yang terdiri atas k pohon.

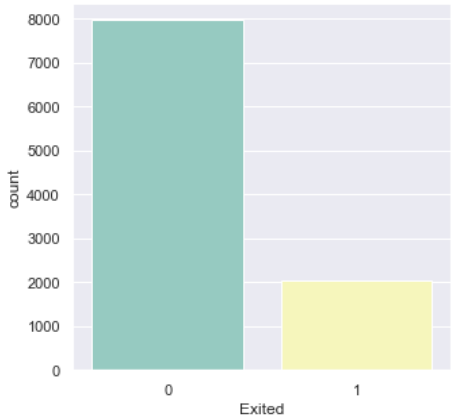
3. Hasil dan Pembahasan

Hal pertama yang dilakukan untuk analisis adalah *mengimport* modul yang diperlukan (*numpy*, *pandas*, *seaborn*, *matplotlib*) serta data *bank customers*.

Number	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Gambar 4. Data Bank Customers

Kemudian dilakukan pelabelan pada data kategorik terutama pada variabel target (*exited)* yaitu 0 untuk retained dan 1 untuk closed.



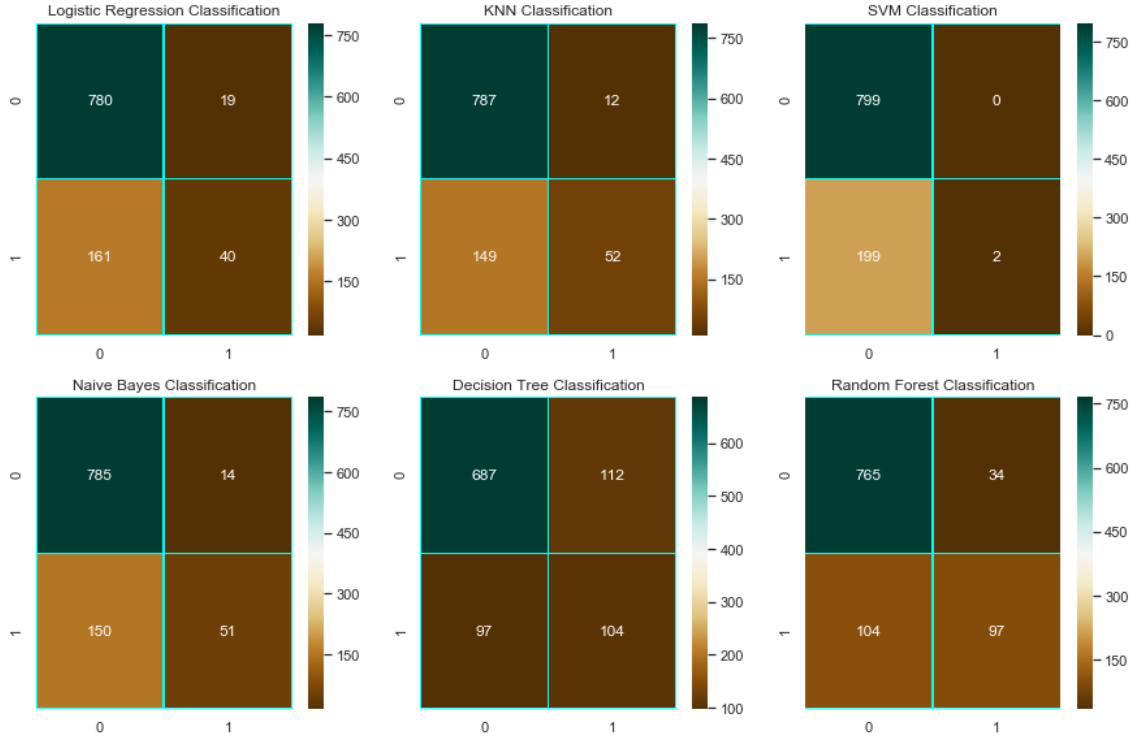
Gambar 5. Visualisasi variabel target

Karena rentamg antar variabel data sangat besar maka perlu dilakukan normalisasi. Setelah dilakukan normalisasi, dataset kemudian dibagi menjadi 2 yaitu data training dan data testing dengan perbandingan 9:1.

	CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
0	0.538	0.0	0.324324	0.2	0.000000	0.000000	1.0	1.0	0.506735
1	0.516	0.0	0.310811	0.1	0.334031	0.000000	0.0	1.0	0.562709
2	0.304	0.0	0.324324	0.8	0.636357	0.666667	1.0	0.0	0.569654
3	0.698	0.0	0.283784	0.1	0.000000	0.333333	0.0	0.0	0.469120
4	1.000	0.0	0.337838	0.2	0.500246	0.000000	1.0	1.0	0.395400

Gambar 6. Data Bank Customers setelah normalisasi

Kemudian dibuatlah model untuk masing masing metode. Setelah model dengan data training terbentuk maka dilakukan pengujian model terhadap data testing. Diperoleh confussion matriks untuk masing-masing metode adalah:



Gambar 7. Visualisasi confussion matrik masing-masing metode

Berdasarkan Gambar 3 maka didapat hasil keakuratan untuk masing-masing metode adalah adalah :

Tabel 1. Keakurasian masing-masing Metode

Metode	Accuracy	Precision	Recall	f1
Regresi Logistik	0,820	0,677	0,199	0,307
K-Nearest Neighboor	0,839	0,813	0,259	0,392
Super Vector Machine	0,801	1,0	0,009	0,020
Naive bayes	0,836	0,785	0,254	0,383
Decision Tree	0,791	0,481	0,517	0,498
Random Forest	0,862	0,740	0,482	0,584

Berdasarkan Tabel 1 didiperoleh nilai akurasi tertinggi adalah 0,862 dengan metode random forest, dan terendah adalah 0,791 dengan metode decision tree. Kemudian untuk precision tertinggi adalah 1,0 dengan metode super vector machine dan terendah adalah 0,481 dengan metode decision tree. Lalu untuk recall teringgi diperoleh 0,517 dengan metode decison tree dan terendah 0,009 dengan metode super vector machine. Kemudian untuk f1 tertinggi adalah 0,584 dengan metode random forest dan terendah adalah super 0,020 dengan metode super vector machine.

4. Simpulan

Berdasarkan pembahasan, maka diperoleh metode terbaik untuk klasifikasi data *bank customers* adalah metode klasifikasi *random forest* hal ini dikarenakan karena ketepatan klasifikasinya yang paling tinggi daripada yang lain dengan nilai akurasi 0,862 atau 86,2%, nilai *precision* 0,740, nilai recall 0,482 dan nilai f1 adalah 0,584.

Daftar Pustaka

McKinney Wes, 2012, *Python for data analysis*, O'Relly

Fractal , (2003), *Comparative Analysis of Classification Techniques*, A Fractal White Paper.

Suyanto, 2018, *Machine learning tingkat dasar dan lanjut*, bandung, informatika

M Reza Faisal, 2019, *Belajar Data Science:Klasifikasi dengan Bahasa Pemrograman R*, Universitas Lambung Mangkurat

Breiman L. 1996. *Bagging Predictors. Machine Learning* 24:123-140.

Breiman L. 2001. *Random Forests. Machine Learning* 45:5-32.

Dewi N.K., Syafitri U.D., & Mulyadi S.Y (2011). *The Application of Random Forest in Driver Analysis.* Forum Statistika dan Komputasi, 16, 35-43.

Nariswari R., & Rafikasari E.F (2019). *Perbandingan Metode Analisis Diskriminan, Neural Network, Diskriminan Kernel, Regresi Logistik, Mars Untuk Data Bangkitan (Kombinasi Varians, Overlap dan Korelasi)*. Open Journal Systems, 13, 1763-1774.