Klasifikasi Kehadiran Pasien: Perbandingan Tiga Metode

Julisa Bana Abraham
Departemen Teknik Elektro dan Teknologi Informasi
Universitas Gadjah Mada
Yogyakarta, Indonesia
julisa.bana.a@mail.ugm.ac.id

***Abstrak*—Paper ini mencoba melakukan klasifikasi terhadap dataset kehadiran pasien di Brazil yang sebelumnya telah melakukan pendaftaran terlebih dahulu, data ini memiliki distribusi kelas yang timpang yaitu sebanyak 79% pasien hadir dan 21% pasien tidak hadir. Klasifikasi dilakukan menggunakan tiga metode yaitu *Naïve Bayes Classifier, Decision Tree, dan Support Vector Machine (SVM*). Pada penelitian ini metode *SVM* dianggap yang paling baik dalam menangani pemodelan klasifikasi untuk dataset ini yang merupakan dataset yang *imbalanced* atau timpang distribusi kelasnya, walaupun memiliki waktu komputasi yang sangat lama, SVM unggul dalam *accuracy* dan *average precision* dibandingkan dengan kedua metode lainnya**

Keywords—Bayesian Classifier, ID3, SVM, Imbalanced Data, K-fold Cross Validation

I. PENDAHULUAN

Klasifikasi merupakan masalah yang sering diselesaikan menggunakan metode *machine learning*. Masalah ini termasuk kedalam jenis *supervised learning* yang memerlukan data latih yang terlabeli dengan benar dalam melakukan pemodelannya, namun banyak dataset yang memiliki distribusi kelas yang tidak berimbang sehingga *accuracy* yang umum menjadi *metrics* untuk performa pemodelan dianggap tidak representatif [1] sehingga diperlukan *metrics* lain untuk dataset seperti ini.

Paper ini mencoba melakukan klasifikasi terhadap dataset kehadiran pasien di Brazil yang sebelumnya telah melakukan pendaftaran terlebih dahulu, data ini memiliki distribusi kelas yang timpang yaitu sebanyak 79% pasien hadir dan 21% pasien tidak hadir. Klasifikasi dilakukan menggunakan tiga metode yaitu *Naïve Bayes Classifier, Decision Tree, dan Support Vector Machine (SVM*).

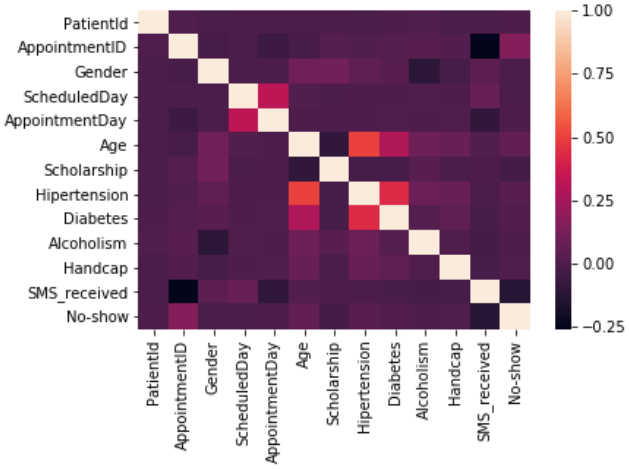
II. MATERIAL DAN METODE

Semua model diimplementasi menggunakan *personal computer* yang berspesifikasi Intel core i7 4700MQ, RAM 16GB.

A. Dataset.

Dataset diperoleh dari Kaggle yang berjudul Medical Appointment No Shows yang berisi kehadiran pasien (datang atau tidak) disertai dengan atribut-atribut yang menyertainya. Dataset ini memiliki 110.527 data kehadiran pasien yang disertai 13 atribut dengan distribusi kelas yang timpang yaitu 79% hadir dan 21% tidak hadir.

Pada penelitian ini seleksi fitur dilakukan berdasarkan korelasi antara kelas yaitu No-show dengan atribut lain. Heatmap dari korelasi ini ditampilkan pada Gambar 1.



Gambar 1. *Heatmap* korelasi antar atribut

Berdasar pada seleksi fitur menggunakan korelasi, penelitian ini dilakukan menggunakan atribut AppointmentID, Age, Diabetes, Hipertension, Scholarship dan SMS_received untuk pemodelannya.

B. Bayesian Classifier

Bayesian Classifier adalah sebuah meode klasifikasi statistik. Classifier ini membantu untuk menghitung probabilitas dari sample yang diberikan, sebagai contoh A termasuk kelas C, inilah yang disebut *membership probability. Bayesian classifier* dikembangkan dari teorema Bayes. Karakteristik dari classifier ini adalah untuk sebuah atribut A dan efek nilai dari A untuk sebuah kelas N adalah independen dari nilai atribut lain yang dimiliki oleh N. Karakteristik ini dinamakan dengan *class conditional independence* [2].

Untuk menghitung P(H|A) probabilitas dari sebuah hipotesis untuk sebuah observasi A:

P(H): probabilitas prior dari hipotesis H

P(A): probabilitas dari sampel data yang terobservasi

P(A|H): probabilitas dari observasi A, pada hipotesis H

Lalu probabilitas posteori P(H|A) dapat dihitung menggunakan teorema bayes:

$$P(H|A) = (P(A|H) \,.\, P(H))/P(A) \quad (1)$$

Pada paper ini metode *bayesian classifer* diimplementasikan menggunakan bahasa Python 3.5 dengan modul scikit-learn.

C. *Decision Tree*

Decision Tree merupakan metode yang banyak digunakan untuk masalah klasifikai. Metode ini berusaha menemukan model klasifikasi yang tahan terhadap *noise*. Pada penelitian ini digunakan metode decision tree ID3. Metode ID3 membangun model klasifikasi yang dinamakan dengan *decision tree* dari atribut yang dianggap paling penting sampai yang dianggap tidak relevan menggunakan suatu *metrics* statistik yaitu *information gain* [3]. Pada paper metode ID3 akan diimplementasikan menggunakan bahasa Python 3.5 dengan modul scikit-learn.

D. *Support Vector Machine*

Support Vector Machine (SVM) adalah sebuah metode *supervised learning* yang dapat digunakan untuk masalah klasifikasi maupun regresi. Prinsip algoritma ini adalah untuk memisahkan kelas di dalam n dimensi atribut dengan suatu *hyperplane*. Untuk penelitian ini digunakan SVM yang menggunakan kernel *radial basis function* (RBF) yang populer digunakan untuk masalah klasifikasi menggunakan SVM [4]. Pemodelan SVM juga dilakukan menggunakan bahasa Python 3.5 dengan modul scikit-learn.

E. *Validation and Metrics*

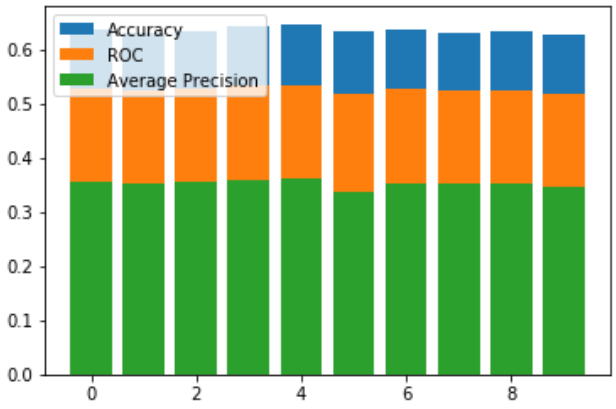
Karena data yang digunakan untuk pemodelan cukup besar maka penelitian ini menggunakan *kfold cross validation* dengan k = 10 untuk mengurangi bias dalam *metrics* yang dihasilkan. Untuk *metrics* pada penelitian ini menggunakan *accuracy* saja dianggap tidak representatif dengan hasil pemodelan karena adanya ketimpangan distribusi kelas. Penelitian ini mencoba untuk menggunakan *Area Under the Curve* (AUC) dari *Receiver Operating Characteristic* (ROC) dan *Average Precision atau Precision Recall curve* sebagai *metrics*, karena biasa digunakan sebagai metrics untuk pemodelan data yang *imbalanced* [5]

III. HASIL DAN DISKUSI

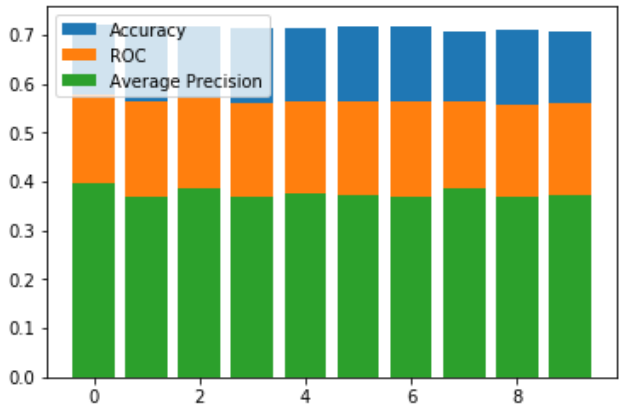
Penelitian ini dievaluasi tiga metode klasifikasi yaitu Bayesian Classifier, Decision Tree, dan SVM menggunakan tiga metrics yaitu *accuracy, precision, dan average precision* yang dilakukan menggunakan *kfold cross validation* dengan k = 10. Hasil rata-rata dari *metrics* dapat dilihat pada Tabel 1. Sedangkan Gambar 2, 3, dan 4 menunjukan *metrics* untuk setiap *fold*.

	Accuracy	**ROC**	**Average Precision**
Bayes	0.6364	0.52646	0.3526
ID3	0.7146	0.5644	0.3761
SVM	0.8023	0.5137	0.5005

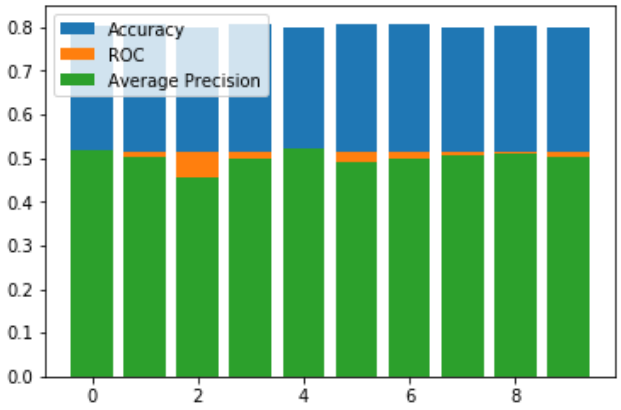
Tabel 1. Hasil rata-rata dari *tenfold cross validation*



Gambar 2. Hasil tenfold cross validation dari bayesian classifier



Gambar 3. Hasil tenfold cross validation dari ID3 classifier



Gambar 4. Hasil tenfold cross validation dari SVM classifier

Ketiga metode memiliki hasil yang tidak jauh berbeda untuk hasil rata-rata ROC, dengan ID3 memiliki ROC rata-rata terbaik dari ketiga model, namun untuk *metrics accuracy* dan *Average Precision* SVM unggul diantara ketiganya.

IV. KESIMPULAN

Pada penelitian ini metode *Support Vector Machine* dianggap yang paling baik dalam menangani pemodelan klasifikasi untuk dataset ini yang merupakan dataset yang *imbalanced* atau timpang distribusi kelasnya, walaupun memiliki waktu komputasi yang sangat lama dibanding kedua metode lainnya.

Masalah pemodelan pada data yang *imbalanced* dapat diatasi menggunakan *metrics* seperti *ROC, average precision*, atau *f1 score*. Namun penelitian ini menitik beratkan pada *metrics average precision* karena lebih representatif untuk mengetahui performa pemodelan untuk dataset *imbalanced* [6] sehingga SVM adalah yang paling tepat untuk dataset ini.

REFERENSI

[1] M. Imran, A. M. Mahmood and A. A. M. Qyser, "An empirical experimental evaluation on imbalanced data sets with varied imbalance ratio," International Conference on Computing and Communication Technologies, Hyderabad, 2014, pp. 1-7.

[2] H. Walia, A. Rana and V. Kansal, "A Naïve Bayes Approach for working on Gurmukhi Word Sense Disambiguation," 2017 6th Optimization (Trends and Future Directions) (ICRITO), Noida, 2017, pp. 432-435

[3] Suyanto, "Data Mining Untuk Klasifikasi dan Klasterisasi data," Penerbit Informatika, February 2017.

[4] Chang, Yin-Wen; Hsieh, Cho-Jui; Chang, Kai-Wei; Ringgaard, Michael; Lin, Chih-Jen (2010). "Training and testing low-degree polynomial data mappings via linear SVM". J. Machine Learning Research. 11: 1471–1490.

[5] Jeni, Laszlo, Cohn, Jeffrey F, & De la Torre, Fernando. (2013). Facing imbalanced data recommendations for the use of performance metrics. Proceedings of the International Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland

[6] Davis, Jesse, Goadrich, Mark. (2006) "The Relationship Between Precision-Recall and ROC Curves". Proceedings of the 23 rd International Con- ference on Machine Learning, Pittsburgh