ISSN 2085-4552

Klasifikasi Diabetes Menggunakan Model Pembelajaran Ensemble Blending

Vinnia Kemala Putri[1], Felix Indra Kurniadi[2]
[1] Fakultas Ilmu Komputer, Universitas Indonesia, Depok, Indonesia
[2] School of Engineering and Technology: Informatics Engineering, Tanri Abeng University, Jakarta, Indonesia
vinnia.kemala51@ui.ac.id
felixindra@tau.ac.id

Diterima 15 Februari 2018
Disetujui 8 Juni 2018

***Abstract*—Diabetes mellitus is one of the deadliest disease and it is increasing in occurrence through the world. This can be prevented by conducting early diagnosis and treatment. However, in developing countries, less than half of people with diabetes are diagnosed correctly which lead to lose of human lives. In this Big Data era, medical databases have enormous quantities of data about their patients. But this medical data may contain noise and a lot of useless information which may mislead the expert in making a decision for medical diagnosis. Data mining is a technique to that is very effective for medical applications for identifying patterns and extracting useful information for databases. This paper proposed a data mining approach using an ensemble blending method to tackle a diabetes prediction problem in Pima Indian Diabetes Dataset. We proposed a blending ensemble classifier approach using a combination of Decision Tree and Logistic Regression as base classifiers, and Support Vector Machine as a top blender classifier. Our approach reached accuracy of 81% and F1-score of 0.81 proves to be higher when compared with basic classifier without combination.**

***Index Terms*—diabetes, ensemble, data mining**

I. PENDAHULUAN

Diabetes merupakan suatu penyakit yang terjadi akibat kurangnya insulin atau ketidakmampuan pancreas dalam tubuh untuk memproduksi hormon insulin. Insulin berguna untuk mengubah glukosa dalam tubuh menjadi glikogen dan trigliserida. Kedua zat ini merupakan sumber energi yang disimpan dalam tubuh. Kekurangan insulin dapat mengakibatkan menumpuknya kadar glukosa dalam darah. Penumpukkan tersebut dapat menyebabkan komplikasi pada ginjal dan beberapa masalah lainnya [1].

Diagnosis medis merupakan hal yang penting dan krusial. Kesalahan diagnosis yang dilakukan oleh dokter dapat mengakibatkan malpraktek dan berujung pada kematian pasien. Hasil analisis data yang diambil dari pasien dan kemampuan dokter untuk mengambil keputusan sangatlah mempengaruhi hasil diagnosis medis. Tahap-tahap yang dilakukan sangatlah panjang dan membutuhkan waktu dan tenaga yang sangat banyak. Oleh karena itu tindak medis tersebut harus dilakukan secara tepat dan efisien.

Sayangnya, di beberapa negara berkembang ketersediaan dokter ahli dalam diabetes sangatlah terbatas dan dapat mengakibatkan kesalahan diagnosis. Hal ini dapat membuang-buang waktu, tenaga, dan bahkan nyawa pasien. Maka dibutuhkanlah suatu metode yang dapat membantu para ahli kesehatan agar dapat memberikan diagnosis yang lebih tepat dan meningkatkan kesejahteraan masyarakat.

Untuk meningkatkan keakuratan dan ketepatan dalam memberikan diagnosis, penting untuk menganalisis pola data medis pasien. Dengan munculnya era *Big Data* dan adanya algortima yang cepat dan efisien untuk menganilis data, memahami suatu pola data medis adalah suatu hal yang mungkin untuk dilakukan. Namun, dengan ketersediaan data yang sangat besar seringkali data yang didapat dari basis data rumah sakit memiliki banyak derau dan informasi yang tidak dibutuhkan. Hal ini dapat menyesatkan para ahli dalam menganalisis data dan memberikan keputusan.

Penambangan data merupakan suatu teknik pengolahan dan penarikan informasi penting dari sejumlah data yang besar. Teknik ini sering digunakan untuk membantu penelitian terkait masalah kesehatan, yang mana digunakan untuk memprediksi, mendiagnosis dan memberikan solusi perawatan terhadap suatu penyakit [2]. Penambangan data juga telah banyak digunakan untuk mengatasi masalah klasifikasi pada diabetes seperti yang dilakukan oleh M. Panwar et al. menggunakan pendekatan penambangan data dan *K-Nearest Neighbor* (KNN) untuk melakukan klasifikasi *binary* pada *Pima Indian* diabetes [3]. T. Jayalakshmi and A. Santhakumaran melalukan praproses pada *missing value* dan klasifikasi diabetes menggunakan Neural Network (NN) [4]. L. Chang-Shing dan W. Mei-Hui mengusulkan *semantic decision support* untuk membantu pengambilan keputusan diagnosis diabetes dengan menggunakan *Fuzzy Diabetes Ontology* (FDO)

[5]. R. Vaishali, et al. mengusulkan algoritma genetika (GA) untuk fitur seleksi dan *Multi Objective Evolutionary* (MOE) *Fuzzy* untuk klasifikasi diabetes [6]. M. Aakanksha, et al. menggunakan Principal Component Analysis (PCA) untuk mereduksi dimensi dan GA untuk meningkatkan kemampuan NN [7]. E.K. Hashi mengusulkan penggunaan *Decision Tree* C4.5 dan KNN sebagai *supervised* klasifier [8].

Salah satu metode pembelajaran mesin yang kuat adalah model *ensemble*. *Ensemble* mengkombinasikan beberapa model pembelajaran mesin yang sederhana seperti *Naïve Bayes, Decision Tree, Logistic Regression* dan lainnya, menjadi satu kesatuan model yang memiliki performa setingkat model pembelajaran mesin yang kuat. Beberapa teknik *ensemble* yang popular adalah *bagging* [9] dan *boosting* [10].

Ensemble sering digunakan dalam penambangan data dan pembelajaran mesin dikarenakan model tersebut sederhana, mudah diaplikasikan dan tahan terhadap bias dan variansi. D.A. Davis et al. [11] menggunakan *collaborative filtering* untuk melakukan prediksi jenis penyakit. *Ensemble* sering sekali digunakan dalam pemodelan sistem rekomendasi. M. Jahrer et al. [12] menggabungkan beberapa algoritma *collaborative filtering* seperti SVD, *Neighborhood based approaches, Restricted Boltzmann Machine* (RBM), *Asymmetric Factor Model* dan *Global Effect* untuk membuat sistem rekomendasi pada dataset Netflix. R. M. Bell et al. [13] menggunakan teknik *ensemble blending* untuk menggabungkan KNN, *factorization*, RBM, dan *Asymmetric Factor Model*. J. Sill et al. [14] mengusulkan teknik *stacking* dengan pembobotan. *Stacking* adalah salah satu teknik *ensemble* yang mana hasil prediksi pada kumpulan model tingkatl pertama, diberikan ke model tingkat kedua sebagai input. Namun, penelitian diabetes menggunakan model *ensemble* masih sangatlah jarang.

Penelitian ini mengusulkan metode *ensemble blending* dengan mengkombinasikan *Decision Tree* (DT), *Logistic Regression* (LogReg), dan *Support Vector Machine* (SVM) untuk melakukan klasifikasi binari pada data diabetes mellitus. Penelitian ini hanya fokus pada pemodelan menggunakan metode *ensemble blending* untuk meningkatkan akurasi pada proses pengklasifikasian, tidak pada prapengolahan data seperti penghilangan derau atau data tidak berhubungan.

II. EKSPLORASI DATASET

Dataset yang dipakai pada penelitian ini adalah dataset yang berasal dari *National Institute of Diabetes and Digestive and Kidney Diseases* dan dapat diakses secara publik di UCI Machine Learning Repository: Pima Indians Diabetes Dataset [15]. Dataset ini memiliki infornasi tentang 768 pasien wanita dengan 8 diagnosis kondisi medis yang berbeda-beda. Tujuan dari penelitian ini adalah untuk memprediksi apakah seorang pasien terkena diabetes atau tidak.

A. *Ringkasan Dataset*

Dataset terdiri dari 9 kolom sebagai atribut yaitu:

- Pregnancies: Jumlah berapa kali pasien pernah mengandung.
- Glucose: Konsentrasi plasma glukosa selama 2 jam pada tes oral konsentrasi glukosa.
- BloodPressure: Tekanan darah diastole (mm Hg).
- SkinThickness: Tebal lapisan kulit pada bagian triceps (mm).
- Insulin: Kadar 2 jam serum insulin (mu U/ml).
- BMI: Indeks massa tubuh (kg/m^2)
- DiabetesPedigreeFunction: Hubungan genetik pasien dengan keluarga yang menderita diabetes.
- Age: Umur dari pasien (tahun).
- Outcome: Terdiri dari 2 kelas, 0 (negatif diabetes) dan 1 (positif diabetes). Atribut yang akan diprediksi.

III. METODOLOGI

Bagian ini memamparkan tentang langkah-langkah penelitian yang dilakukan. Metode yang dilakukan adalah prapengolahan data, pelatihan model dan evaluasi model.

A. *Prapengolahan Data*

Dikarenakan banyaknya derau pada dataset, prapengolahan data perlu dilakukan untuk meningkatkan akurasi dari penambangan data sebelum mengimplementasikan algoritma pembelajaran mesin pada dataset yang digunakan. Beberapa langkah yang dilakukan yaitu:

- Mengolah *missing values*. Pada dataset ini, selain pada atribut *Outcome*, banyak terdapat angka 0 yang menandakan data tidak tertangkap dengan baik. Angka-angka 0 tersebut terdapat pada atribut Glucose, BloodPressure, SkinThickness, Insulin dan BMI. Angka-angka 0 tersebut diganti menjadi NA dan kemudian diisi dengan menggunakan KNN.
- Melakukan *oversampling* untuk mengatasi ketidakseimbangan jumlah data pada tiap kelas.
- Melakukan normalisasi *center* dan *scale* pada data untuk memperkecil bias dan variansi pada

hasil eksperimen dan mempercepat waktu komputasi.

B. *Metode Ensemble - Blending*

Dalam mengatasi masalah klasifikasi, pendekatan yang paling sering digunakan adalah melatih beberapa model klasifier dan memilih satu yang memberikan hasil terbaik dari *cross validation*. Namun, tiap-tiap memiliki variansi dan bias yang berbeda-beda. Tiap-tiap akan konvergen dan gagal di tempat yang berbeda.

Salah satu metode pembelajaran mesin yang kuat adalah model *ensemble*. *Ensemble* mengkombinasikan beberapa model klasifier dengan tujuan dapat menggabungkan kelebihan dari tiap-tiap model.

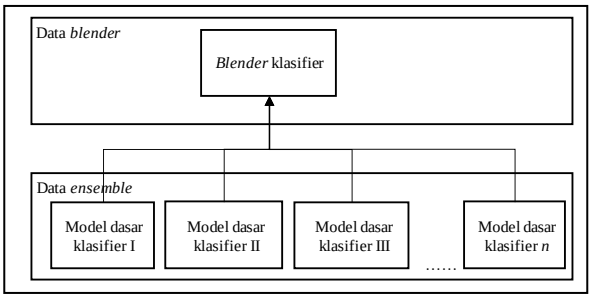
Syarat utama dalam pengaplikasian *ensemble* adalah model-model dasar klasifier harus saling independen dengan tujuan *error rate* dari tiap-tiap model tidak saling berkorelasi. Selain itu, performa tiap-tiap model dasar harus lebih baik dari sebuah klasifier dengan *random guessing* (*error* < 0,5). Model-model dasar dipilih bukan dari tingginya akurasi tetapi dari kesederhanaannya.

Blending adalah salah satu dari teknik *ensemble*. *Blending* pertama kali diperkenalkan oleh pemenang kompetisi Netflix 2007 [13]. Secara umum, cara kerja *blending* mirip dengan *stacked generalization* [16]. Ide dasar dari *blending* adalah menggunakan beberapa klasifier sebagai model dasar kemudian menggunakan sebuah *blender* klasifier untuk mengkombinasikan prediksi-prediksi dari model-model dasar.

Beberapa langkah dari *blending* yaitu:

- Membagi data pelatihan menjadi 2 bagian: data *ensemble* dan data *blender*.
- Melakukan pelatihan model-model dasar dengan menggunakan data *ensemble*.
- Melakukan prediksi untuk data *blender* dan data pengujian dengan menggunakan model-model dasar yang telah dilatih sebelumnya.
- Hasil prediksi oleh model-model dasar untuk data *blender*, dijadikan atribut baru untuk data *blender*. Demikian juga hasil prediksi oleh model-model dasar untuk data pengujian, dijadikan atribut baru untuk data pengujian.
- Melatih model akhir *blender* klasifier dengan menggunakan data *blender* yang telah ditambah atributnya.
- Melakukan prediksi dengan menggunakan model akhir *blender* klasifier untuk data pengujian yang telah ditambah atributnya.

Gambar 1 mengilustrasikan ide dasar metode *blending*.



Gambar 1. Ide dasar metode *ensemble blending*

C. *Evaluasi Model*

Hasil penelitian dievaluasi dengan mengunakan akurasi dan F1-score.

Metode lain untuk mengevaluasi hasil klasifikasi dari model yang berbeda-beda dapat menggunakan *confusion matrix*. Dalam klasfikasi biner, *confusion matrix* memberikan beberapa detail sebagai berikut:

- *True Positive* (TP): Data pada kelas positif diklasifikasikan pada kelas positif.
- *True Negative* (TN): Data pada kelas negatiff diklasifikasikan pada kelas negatif.
- *False Positive* (FP): Data pada kelas negatif diklasifikasikan pada kelas positif.
- *False Negative* (FN): Data pada kelas positif diklasifikasikan pada kelas negatif.

Akurasi mengindikasikan rasio antara data teridentifikasi dengan tepat dan keseluruhan data pengujian. Persamaan untuk mengukur akurasi adalah sebagai berikut:

$$akurasi = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (1)$$

Sensitifitas mengindikasikan rasio antara TP dengan keseluruhan data aktual kelas positif. Persamaan untuk mengukur sensitifitas adalah sebagai berikut:

$$sensitifitas = \frac{TP}{FN + TP} \quad (2)$$

Presisi mengindikasikan rasio antara TP dengan keseluruhan data prediksi kelas positif. Persamaan untuk mengukur sensitifitas adalah sebagai berikut:

$$presisi = \frac{TP}{FP + TP} \quad (3)$$

F1-score mengindikasikan ratio rata-rata antara sensitifitas dan presisi. Persamaan untuk mengukur F1-score adalah sebagai berikut:

$$F1 - score = \frac{2 \times presisi \times sensitifitas}{presisi + sensitifitas} \quad (4)$$

IV. EKSPERIMEN DAN HASIL

Bagian ini menjelaskan tentang eksperimen yang dilakukan untuk memprediksi kondisi pasien terkena diabetes atau tidak. Eksperimen dilakukan dengan menggunakan bantuan *tool* R studio.

Dataset awal berjumlah 768 pasien dengan 500 pasien tidak terkena diabetes (kelas 0) dan 268 pasien terkena diabetes (kelas 1). Dikarenakan terdapat ketidakseimbangan jumlah data pada tiap kelas, maka dilakukan *oversampling* data pada kelas dengan jumlah data lebih kecil, yaitu kelas 1. Setelah dilakukan *oversampling*, jumlah data pada kelas 1 menjadi 500. Dataset akhir yang digunakan berjumlah 1000 pasien.

Proporsi data pelatihan dan pengujian adalah 70% (700 pasien) dan 30% (300 pasien) secara berurutan. Metode validasi yang digunakan adalah *k-fold cross validation* dengan nilai k = 10.

A. *Eksperimen I – Pemilihan Model*

Pada eksperimen I melakukan beberapa klasfikasi dengan menggunakan algoritma pembelajaran mesin yang berbeda-beda guna menentukan model dasar untuk model *ensemble* nantinya. Algoritma yang digunakan adalah *decision tree, k-nearest neighbor, logistic regression,* dan SVM. Eksperimen menggunakan keseluruhan data pelatihan dan data pengujian. Tabel 1 memaparkan hasil evaluasi dari klasifikasi data diabetes Pima Indian pada data pengujian.

Tabel 1. Perbandingan hasil evaluasi

Algoritma	Akurasi	F1-Score
Decision tree	74,33%	0,74
k-nearest neighbor	73,33%	0,73
Logistic regression	75,33%	0,76
SVM (kernel linear)	73,67%	0,73
SVM (kernel radial)	74,00%	0,74

Pada tabel I dapat terlihat bahwa model dengan nilai akurasi dan F1-score terbaik adalah *logistic regression* dengan akurasi sebesar 75,33% dan F1 score 0,76.

B. *Eksperimen II – Model Ensemble*

Pada tahap eksperimen ini, pemilihan model dasar pembelajaran mesin didasari pada hasil eskperimen I. Model yang dipilih adalah *decision tree, logistic regression* dan SVM dengan kernel radial. Pembagian data *ensemble* dan data *blender* masing-masing adalah 50% dari data pelatihan.

Tahap ini melatih tiga kombinasi model yang berbeda, yaitu:

- Model I; dengan model dasar adalah *decision tree* dan *logistic regression*, dan *blender* adalah SVM dengan kernel radial.
- Model II; dengan model dasar adalah *decision tree* dan SVM dengan kernel radial, dan *blender* adalah *logistic regression*.
- Model III; dengan model dasar adalah *logistic regression* dan SVM dengan kernel radial, dan *blender* adalah *decision tree*.

Tabel 2. Perbandingan hasil evaluasi

Model	Akurasi	F1-Score
I. Model dasar: DT+LogReg *Blender*: SVM (kernel radial)	81,00%	0,81
II. Model dasar: DT+ SVM (kernel radial) *Blender*: LogReg	80,33%	0,81
III. Model dasar: LogReg+ SVM (kernel radial) *Blender*: DT	79,67%	0.79

Pada tabel 2 dapat dilihat bahwa model *ensemble* memberikan peningkatan yang signifikan daripada model tanpa dilakukan *ensemble* seperti yang tertera pada tabel 1. Tabel 3-5 memaparkan *confusion matrix* untuk masing-masing model.

Tabel 3. *Confusion Matrix* Model I *Blender* SVM

n=300		Aktual	
		0	**1**
Prediksi	**0**	TN=124	FN=18
	1	FP=39	TP=119

Tabel 4. *Confusion Matrix* Model II *Blender* LogReg

n=300		Aktual	
		0	**1**
Prediksi	**0**	TN=107	FN=26
	1	FP=33	TP=134

Tabel 5. *Confusion Matrix* Model III *Blender* DT

n=300		Aktual	
		0	**1**
Prediksi	**0**	TN=121	FN=31
	1	FP=30	TP=118

V. SIMPULAN

Penelitian ini mengusulkan penggunaan model *ensemble blending* untuk mengatasi masalah klasifikasi diagnosis diabetes. Dari hasil evaluasi dapat terlihat bahwa metode yang diusulkan memberikan nilai akurasi dan F1-*score* yang lebih tinggi daripada model pembelajaran mesin biasa.

Banyak cara yang dapat dilakukan untuk meningkatkan performa dari model yang diusulkan ini. Salah satunya dengan melakukan seleksi fitur dan reduksi dimensi pada praproses dataset. Diharapkan dengan melakukan praproses data yang lebih baik, dapat mengurangi derau pada data dan otomatis meningkatkan performa dari model.

Dengan adanya penelitian ini diharapkan dapat membantu para ahli medis dalam mengambil keputusan ketika sedang melakukan diagnosis medis kepada pasien. Sehingga, nilai kematian pasien akibat kesalahan diagnosis penyakit diabetes dapat dikurangi. memberikan informasi mengenai kecenderungan.

DAFTAR PUSTAKA

[1] N. Yilmaz, O. Inan, dan MS. Uzer, "New data preparation method based on clustering algorithm for diagnosis systems of heart and diabetes diseases," J. Med Syst vol. 38, 2014 hal 48-59.

[2] Richards, Graeme, et al, "Data mining for indicators of early mortality in a diabetes of clinical records." Artificial intelligence in medicine 22.3, 2001, hal 215-231.

[3] M. Panwar, et al, "K-Nearest Neighbor Based Methodoogy for Accurate Diagnosis of Diabetes Melitus," Sixth International Symposium on Embedded Computing and System Design (ISED), 2016.

[4] T. Jayalakshmi dan A. Santhakumaran, "A novel classification method for diagnosis of diabetes melitus using atificial neural networks." Data storage and Data Engineering (DSDE), 2010 International Conference on. IEEE, 2010.

[5] L. Chang-Shing dan W. Mei-hui. "A fuzzy expert system for diabetes decision support application." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 41.1, 2011, hal 139-153.

[6] R. Vaishali, et al, "Genetic algorithm based feature selection and MOE fuzzy classification algorithm on Pima Diabetes dataset." IEEE International Conference on Computing Networking and Informatics (ICCNI), 2017.

[7] M. Aakanksha, et al, "Diagnosis of Diabetes Mellitus using PCA and Genetically Optimized Neural Network." IEEE International Conference on Computing, Communication and Automation (ICCCA), 2017.

[8] E.K. Hashi, et al, "An expert clinical decision support system to predict disease using classification techniques." IEEE International Conference on Electrical, Computer and Communication Engineering (ECCE), 2017.

[9] L. Breiman," Bagging Predictor." Techinical Report 421, Departement of Statistics, University of California, 1994.

[10] Y. Freund dan R. Schapire," Experiments with a new Boosting Algorithm." In Proceedings of the Thirteenth International Conference on Machine Learning, 1996, hal148-156.

[11] D. A. Davis, N. V. Chawla, N. A. Christakis, dan A.-L. Barabasi."Time to CARE: a collaborative engine for practical disease prediction." Springe, 2009.

[12] M. Jahrer, et al,"Combining Predictions for Accurate Recommender Systems." 2010.

[13] R. M. Bell, Y. Koren, dan C. Volinsky,"The BellKor Solution to the Netflix Prize." 2007.

[14] J. Sill, G. Takacs, L. Mackey, dan D. Lim,"Feature-Weighted Linear Stacking." Arxiv:0911.04602v2, 2009

[15] J.W. Smith, et al. "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus." IEEE Computer Society Press. In Proceedings of the Symposium on Computer Applications and Medical Care, 1988, hal 261-265.

[16] D.H. Wolpert, "Stacked Generalization." Neural Network 5(2), 1992, hal 241-259.