OPTIMASI ALGORITMA NAÏVE BAYES BERBASIS PARTICLE SWARM OPTIMIZATION (PSO) DAN STRATIFIED UNTUK MENINGKATKAN AKURASI PREDIKSI PENYAKIT DIABETES

Asep Muhidin[1)], Muhamad Casdi[2)]

Program Studi Teknik Informatika Fakultas Teknik
Universitas Pelita Bangsa
asep.muhidin@pelitabangsa.ac.id

Disetujui, 25 September 2019

Abstraksi

Diabetes militus adalah penyakit yang menyebabkan jumlah kadar gula dalam darah tak terkontrol karena kurangnya kadar hormon insulin dalam tubuh. Berdasarkan data hasil tes laboratorium diabetes dapat diprediksi dengan data mining yang dapat membantu para tenaga medis. Data mining adalah suatu proses mengidentifikasi data agar menjadi sebuah informasi maupun keputusan. Penelitian ini menggunakan algoritma Naïve Bayes berbasis Particle Swarm Optimization (PSO) dan Stratified. Hasil dari algoritma Naïve Bayes mendapatkan nilai akurasi 75.40% dan nilai AUC 0,829%. Sedangkan hasil dari algoritma Naïve Bayes berbasis Particle Swarm Optimization (PSO) dan Stratified mendapatkan nilai akurasi 90,00% dan nilai AUC 0,926. Dari penelitian ini bahwa algoritma Naïve Bayes berbasis Particle Swarm Optimization (PSO) dan Stratified Mendapatkan nilai akurasi yang lebih tinggi dengan peningkatan 14,60% dalam memprediksi penyekit Diabetes.

Keyword : Diabetes, *Data Mining, Naïve Bayes, PSO, Stratified*

Abstract

Diabetes mellitus is a disease that causes uncontrolled blood sugar levels due to a lack of insulin levels in the body. Based on data, the results of diabetes laboratory tests can be predicted with data mining that can help medical personnel. Data mining is a process of identifying data to become information and decisions. This study uses the Naïve Bayes algorithm based on Particle Swarm Optimization (PSO) and Stratified. The results of the Naïve Bayes algorithm get an accuracy value of 75.40% and an AUC value of 0.829%. Meanwhile, the results of the Naïve Bayes algorithm based on Particle Swarm Optimization (PSO) and Stratified get an accuracy value of 90.00% and an AUC value of 0.926. From this study, the Particle Swarm Optimization (PSO) and Stratified based Naïve Bayes algorithm obtained a higher accuracy value with an increase of 14.60% in predicting diabetes disease.

Keyword : Diabetes, *Data Mining, Naïve Bayes, PSO, Stratified*

1. Pendahuluan

Teknologi Informasi semakin berkembang dari masa ke masa dan mengakibatkan tingkat kebutuhan akan informasi semakin meningkat. Informasi sangat dibutuhkan dalam berbagai bidang salah satunya bidang kesehatan. Keakuratan informasi dalam sebuah keputusan sangat dibutuhkan untuk mendapatkan hasil yang terbaik. Banyaknya informasi dalam dunia kesehatan dapat dijadikan sebuah data untuk mendapatkan hasil yang diinginkan dengan data mining. Data tersebut dapat dijadikan untuk mengambil sebuah keputusan(Wulandari, 2013). Dalam bidang kesehatan data mining dapat digunakan untuk memprediksi suatu penyakit, salah satunya adalah penyakit diabetes. Diabetes militus atau biasa disebut kencing manis adalah penyakit yang menyebabkan jumlah kadar gula dalam darah tak terkontrol karena kurangnya kadar hormon insulin dalam tubuh. Diabetes menyebabkan 1,5 juta kematian pada tahun 2012. Gula darah yang lebih tinggi dari batas maksimum mengakibatkan tambahan 2,2 juta kematian, dengan meningkatkan risiko penyakit kardiovaskular dan lainnya. Empat puluh tiga persen (43%) dari 3,7 juta kematian ini terjadi sebelum usia 70 tahun(Global Report, 2016). Data laboratorium yang belum difungsikan secara efektif bisa digunakan untuk deteksi penyakit diabetes, dalam data mining teknik klasifikasi adalah lebih popular dalam diagnosis medis untuk memprediksi penyakit. Menurut penelitian yang dilakukan(Wibowo, 2015) pada data gangguan motorik dasar anak pengujian dengan hanya menggunakan Naïve Bayes mendapatkan nilai accuracy sebesar 88,67%.

Sedangkan pengujian dengan menggunakan Naïve Bayes classifier berbasis Adaboost sebesar 90.00% Dan pengujian dengan menggunakan algoritma Naive Bayes classifier dengan optimasi PSO sebesar 98.00%. Maka peneliti tertarik untuk menggunakan Algoritma Naive Bayes Berbasis Particle Swarm Optimization (PSO) dan Stratified untuk meningkatkan akurasi prediksi penyakit DiabetesRumusan Masalah Berdasarkan latar belakang diatas, maka dapat diambil rumusan masalah sebagai berikut :

1. Bagaimana memprediksi penyakit diabetes dengan algoritma naïve bayes.
2. Seberapa besar efektivitas PSO dalam meningkatkan performa dari algoritma naïve bayes dalam prediksi penyakit diabetes?
3. Bagaimana perbandingan nilai akurasi antara algoritma naïve bayes dan algoritma naïve bayes yang berbasis PSO?

2. Tinjauan Studi

2.1. Diabetes

Diabetes militus atau biasa disebut kencing manis adalah penyakit yang menyebabkan jumlah kadar gula dalam darah tak terkontrol karena kurangnya kadar hormon insulin dalam tubuh, diabtes adalah penyakit yang berlangsung lama atau kronis serta ditandai dengan kadar gula (glukosa) darah yang tinggi atau diatas nilai normal (Marianti, 2018).

2.2. Algoritma Naive Bayes

Naive Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma mengunakan teorema Bayes dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variabel kelas. Menurut (Han, 2012) algoritma ini mengandalkan sebuah peluang kemungkinan suatu objek.

2.3. Paticele Swam Optimization

Particle Swarm Optimization (PSO) diperkenalkan oleh Dr. Eberhart dan Dr. Kennedy pada tahun 1995, merupakan algoritma optimasi yang meniru proses yang terjadi dalam kehidupan populasi burung (flock of bird) dan ikan (school of fish) dalam bertahan hidup (Sreejini, 2015). Menurut Cholissodin (2016) Kata "particle" menunjukan individu, misalnya seekor burung dalam kawanan burung. Setiap individu atau partikel berprilaku saling terhubung dengan cara menggunakan kecerdasanya (intelligence) sendiri dan juga dipengaruhi prilaku kelompok kolektifnya. Dengan demikian, jika sutu partikel atau seekor burung menemukan jalan yang tepat atau pendek menuju sumber makanan, sisa kelompok yang lain juga akan dapat segera mengikuti jalan tersebut meskipun lokasi mereka jauh dari kelompok tersebut (Cholissodin, 2016).

2.4. Stratifed

Stratified random sampling adalah suatu teknik pengambilan sampel dengan memperhatikan suatu tingkatan (strata) pada elemen populasi. Elemen populasi dibagi menjadi beberapa tingkatan (stratifikasi) berdasarkan karakter yang melekat padanya. Dalam stratified random sampling elemen populasi dikelompokkan pada tingkatan- tingkatan tertentu dengan tujuan pengambilan sampel akan merata pada seluruh tingkatan dan sampel mewakili karakter seluruh elemen populasi yang heterogen (Yuwono, 2018).

3. Desain Penelitian /Metodologi

Penelitian ini menggunakan penelitian eksperimen. Penelitian eksperimen melibatkan penyelidikan perlakuan pada parameter atau variabel tergantung dari penelitinya dan menggunakan tes yang dikendalikan oleh si peneliti itu sendiri (Rinawati, 2017) Dengan metode penelitian sebagai berikut:

1. Pengumpulan data
   Pada tahap ini menentukan data yang akan di proses, mencari data yang tersedia.
2. Data Preprocessing
   Pada tahap ini menjelaskan tentang tahapan awal data mining yang meliputi penentuan atribut dan pembersian data.
3. Metode yang diusulkan
   Pada tahap ini data dianalisis. Setelah data dianalisis lalu diterapkan model-model yang sesuai dengan jenis data.
4. Percobaan dan pengujian metode Pada tahap ini dilakukan percoban meliputi cara pemilihan arsitektur yang tepat dari metode yang diusulkan sehingga didapatkan hasil yang

dapat membuktikan bahwa metode yang digunakan adalah tepat.

5. Evaluasi dan validasi hasil

   Pada tahap ini dilakukan evaluasi dan validasi hasil penerapan terhadap model penelitian yang dilakukan untuk mengetahui tingkat keakurasian model.

3.1. Pengumpulan data

Pada penelitian ini objek yang akan diambil adalah sebuah data public yang diambil dari situs datahub.io dengan judul Pima Indians Diabetes Database (PIDD) yang bersumber dari UCI Machine Learning Repository. Dataset Pima Indians Diabetes Database (PIDD) ini mempunyai 768 data dan mempunyai 8 atribut 1 label atau kelas. Label class dengan hasil Positif yang dinyatakan mengidap penyakit Diabtes.

3.2. Data Preprocessing

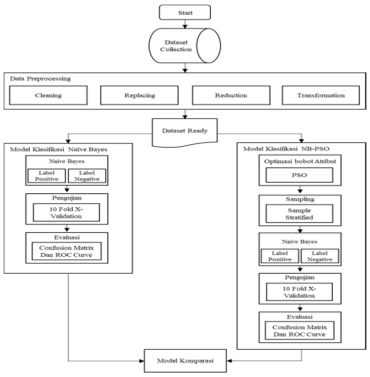
Dataset Pima Indians Diabetes Database (PIDD) ini mempunyai 768 data dan mempunyai 8 atribut 1 label atau kelas. Label class dengan hasil Negative berjumlah 500 data dan Positif dengan jumlah 268 data. Dalam membuat sebuah keputusan atau menentukan akurasi perlu adanya data yang berkualitas untuk itu dilakukan tahap data preprocessing. Data preprocessing digunakan untuk membersihkan data dari missing value, ketidak konsistenan, data tidak lengkap dan noise data. Untuk menghilangkan duplikasi dan inkonsistensi data, maka dilakukan replace missing.

Tabel 1. Jumlah *missing Value*

Atribut	Singka tan	Jumlah Missing Value
Pregnant	Pregnant	111
Plasma Glucose	Glucose	5
Diastolic Blood Pressure	DBP	35
Tricepts Skin Fold Thickness	TSFT	227
Insulin	INS	374
Body Mass	BMI	11
Diabetes pedigree function	DPF	Langkap
Age	Age	Langkap
Class variable	class	Langkap

3.3. Metode yang diusulkan

Metode yang diusulakan pada penelitia ini adalah menggunakan dua pengujian yang pertama menggunakan algoritma *Naive Bayes* dan untuk pengujian yang kedua menggunakan algoritma *Naïve Byes* berbasis *Particle swarm optimization* dan *Stratified.* Berikut ini bentuk gambaran metode algoritma yang akan diuji. Dari kedua pengujian tersebut akan divalidasi dengan *k-fold Cross Validation* dengan jumlah k=10 dan di *evaluasi* oleh *Confusion Matrix* dan *ROC Curve.* Hasil dari kedua pengujian tersebut akan dilakukan model komparasi untuk melihat metode pengujian mana yang mendapatkan nilai yang tebaik. Berikut gambar 1 alur model yang diusulkan :



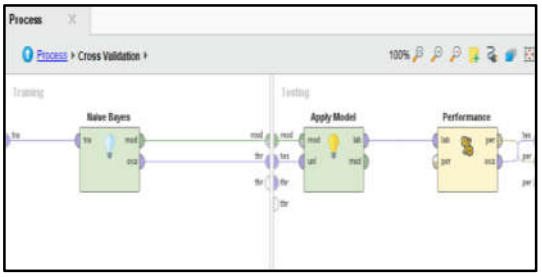
Gambar 1. Alur model yang diusulkan

4. Hasil dan Pembahasan

4.1 Pengujian 1 (Naïve Bayes)

Hasil dari pengujian metode yang dilakukan adalah untuk menentukan nilai accuracy dan AUC. Metode pengujiannya menggunakan cross validation dengan desain modelnya sebagai berikut.

Gambar 2. Model pengujian 1



Gambar 3. Sub Proses *Cross Validation*

Nilai *Accuracy*, dan AUC dapat dihitung dengan menggunakan RapidMiner.

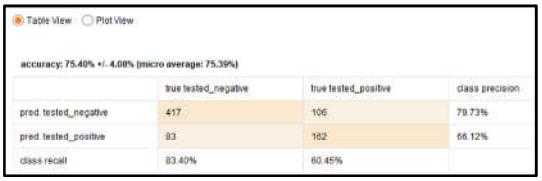


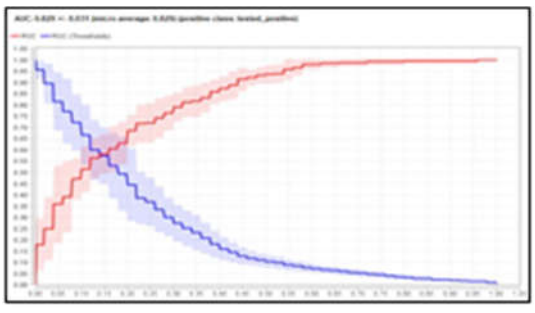
Table View | Plot View

accuracy: 75.40% +/- 4.08% (micro average: 75.39%)

	true tested_negative	true tested_positive	class precision
pred. tested_negative	417	106	79.73%
pred. tested_positive	83	162	66.12%
class recall	83.40%	60.45%	

Gambar 4. Hasil *Accuracy* pengujian 1

Pada gambar 4 menjelaskan hasil pengujian model klasifikasi dengan metode naïve bayes pada confusion matrix didapatkan hasil dari 768 data. 417 data diklasifikasikan Negative sesuai dengan prediksi yang dilakukan dengan metode naïve bayes, lalu 106 data diprediksi Negative tetapi ternyata Positive. Selanjutnya 83 data diprediksi Positive ternyata hasilnya Negative dan 162 data Positive sesuai dengan prediksinya.

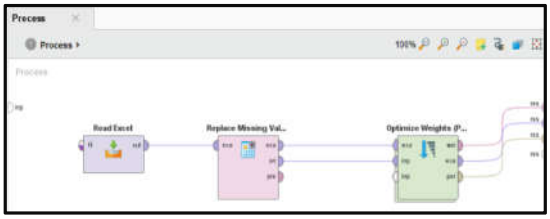


Gambar 5. AUC hasil pengujian 1

Pada gambar grafik 5 menunjukan hasil pengujian algoritma naïve bayes nilai AUC yang diproleh sebesar 0,829 dan masuk dalam klasifikasi *Good classification.*

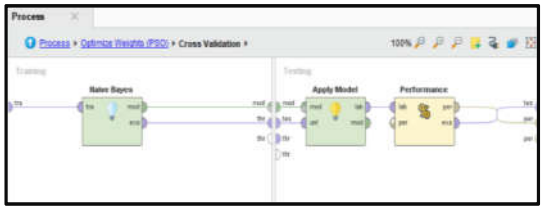
4.2 Pengujian 2 (Naïve Bayes berbasis PSO dan Stratified)

Hasil dari pengujian metode yang dilakukan adalah untuk menentukan nilai accuracy dan AUC. Kemudian untuk membedakan dengan pengujian pertama ialah penerapan fitur algoritma PSO untuk memberikanbobot pada setiap atribut dataset diabetes, pada penelitian ini pemberian bobot pada parameter population size 30-60 dan maximum number of generation 30-60. Metode pengujiannya menggunakan *cross validation* dengan desain modelnya sebagai berikut :



Gambar 6. Model Pengujian 2

Gambar 7. Sub Proses PSO



Gambar 8. Sub Proses *Cross Validation*

Tabel 2. Hasil pengujian 2

NB-PSO		Parameter PSO
Akurasi	**AUC**	
88	0,921	Posize=30, Max num=30
88	0,917	Posize=30, Max num =40
88	0,938	Posize=30, Max num =50
90	0,926	Posize=30, Max num =60
88	0,896	Posize=40, Max num =30
87	0,911	Posize=40, Max num =40
88	0,881	Posize=40, Max num =50
87	0,915	Posize=40, Max of num 60
86	0,878	Posize=50, Max num =30
88	0,919	Posize=50, Max num =40
88	0,919	Posize=50, Max num =50
88	0,919	Posize=50, Max num =60
86	0,907	Posize=60, Max num =30
86	0,938	Posize=60, Max num =40
87	0,922	Posize=60, Max num =50
89	0,911	Posize=60, Max num =60

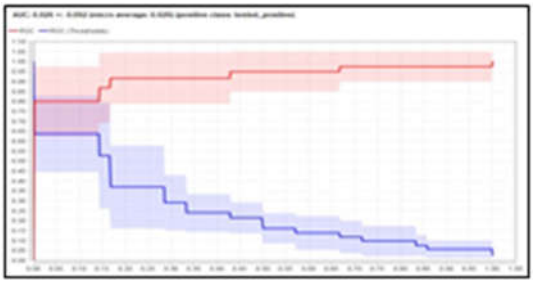
Berdasarkan tabel diatas hasil pengujian yang diperoleh akurasi tertinggi terjadi pada saat population size bernilai 30 dan maximum number of generation bernilai 60. Artinya hasil akurasi sudah maximal, jika

dilakukan percobaan lagi dengan population size dan maximum number of generation lebih besar dari 60 maka akan terjadi overfiting (kelebihan iterasi) yang berdampak pada meningkatnya waktu eksekusi sedangkan hasil akurasi tidak meningkat. Dengan demikian diketahui bahwa algoritma naive bayes berbasis PSO menghasilkan nilai akurasi terbaik pada saat population size bernilai 30 dan maximum number of generation bernilai 60 yaitu nilai akurasi 90.00% dan AUC 0,926

Gambar 9. Hasil *Accuracy* pengujian 2

Pada gambar 8. Menjelaskan hasil pengujian model klasifikasi dengan metode naïve bayes berbasis PSO pada confusion matrix didapatkan hasil dari data sampling bertingkat sebanyak 100 data. 62 diklasifikasikan Negative sesuai dengan prediksinya, lalu 7 data diprediksi Negative tetapi ternyata Positive. Selanjutnya 3 data diprediksi Positive ternyata hasilnya Negative dan 28 data Positive sesuai dengan prediksinya.

Berdasarkan hasil dari penelitian yang telah dilakukan peneliti memberikan beberapa saran yang dapat digunakan pada penelitian kedepanya sebagai berikut



Gambar 10. AUC hasil pengujian 2

Pada gambar 9. Menunjukan hasil pengujian algoritma *naïve bayes* berbasis PSO nilai AUC diproleh sebesar 0,926 dan masuk dalam klasifikasi *Excellent classification*.

5. Kesimpulan

Berdasarkan seluruh hasil tahapan penelitian yang telah dilakukan peneliti mulai dari awal hingga proses pengujian maka dapat disimpulkan bahwa:

1. Hasil penelitian yang dilakukan metode algoritma naïve bayes mendapatkan hasil Accuracy 75,4% dan AUC 0,829 sedangkan metode algoritma naïve bayes berbasis PSO dan Stratified mendapatkan hasil Accuracy 90,0% dan AUC 0,926. Maka nilai Accuracy tertinggi didapatkan dari metode algoritma naïve bayes berbasis PSO dan Stratified.
2. Berdasarkan penelitian yang sudah dilakukan metode algoritma Naïve Bayes berbasis Particle Swarm Optimization (PSO) dan Stratified nilai Accuracy mengalami peningkatan sebesar 14,60%. Metode algoritma PSO dan Stratified efektif dalam meningkatan performa algoritna Naïve Bayes.

Daftar Pustaka

Cholissodin and E. Riyandani, "(Teori & Case Study)," 2016.

D. R. Wulandari and Y. J. Sugiri, "Diabetes Melitus dan Permasalahannya pada Infeksi Tuberkulosis," vol. 33, no. 2, pp. 126–134, 2013.

D. Yuwono, "Stratified Random Sampling: Pengertian dan Konsep Dasar," *Statmat.id*, 2018. [Online]. Available: https://statmat.id/stratified- random-sampling-adalah/.

J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

K. S. Sreejini and V. K. Govindan, "Improved multiscale matched filter for retina vessel segmentation using PSO algorithm," *Egypt. Informatics J.*, 2015.

K. Wibowo, Sfenrianto, and K. Nainggolan, "KLASIFIKASI GANGGUAN MOTORIK KASAR ANAK MENGGUNAKAN NAIVE BAYES SERTA OPTIMASI DENGAN PSO DAN ADABOOST," vol.

1, no. 1, pp. 1–10, 2015.

Marianti, "Diabetes,"*alodokter.com*, 2018. [Online]. Available: https://www.alodokter.com/diabe tes.

R. Rinawati, "Penentuan Penilaian Kredit Menggunakan Metode Naive Bayes Berbasis Particle Swarm Optimization," *J- SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 1, no. 1, p. 48, 2017.

W. Global Report, *Global Report on Diabetes*. 2016.