Prediksi Pima Indians Diabetes Database Dengan Ensemble Adaboost Dan Bagging

Rousyati[1], Amin Nur Rais [2], Eka Rahmawati[3], Richky Faizal Amir[4]

[1]Prodi Sistem Informasi Kampus Kota Tegal, Universitas Bina Sarana Informatika
[2,4]Prodi Teknologi Komputer, Universitas Bina Sarana Informatika
[3]Prodi Sistem Informasi Kampus Kota Surakarta, Universitas Bina Sarana Informatika

** Corresponding Author.E-mail: rousyati.rou@bsi.ac.id*

Abstrak

Teknologi informasi dan komunikasi dapat digunakan oleh para pakar ataupun dokter untuk menafsirkan tentang penyakit dalam waktu yang cepat dan akurat. Salah satu penerapan teknologi informasi di dunia Kesehatan dapat digunakan untuk memprediksi penyakit diabetes. Penelitian ini akan memprediksi pima indians diabetes database dengan ensemble adaboost dan bagging untuk menghasilkan akurasi yang lebih tinggi dalam mendeteksi penyakit diabetes. Data mining adalah suatu teknologi yang dapat digunakan untuk membantu perusahaan dalam mencari informasi yang dapat digunakan dari data yang dimiliki. Penggunaan data mining di implementasikan untuk prediksi, untuk mempredisi apa yang terjadi di masa yang akan datang. Dalam penelitian ini membandingkan penggunaan ensemble adaboost dan bagging terhadap algoritma klasifikasi Naïve Bayes, SVM, dan Decision Tree untuk menghasilkan nilai akurasi dan presisi terbaik dari dataset Pima Indians Diabetes Database. Penelitian ini telah dapat diketahui bahwa nilai akurasi terbaik menggunakan algoritma SVM dengan penggabungan ensemble bagging meski perubahan nilai akurasinya tidak mengalami kenaikan yang signifikan sebesar 77,47%. Namun, pada pengujian presisi dihasilkan penggunaan naïve bayes lebih baik tanpa menggunakan ensemble baik adaboost maupun bagging dengan nilai 80,23%.

Keywords: Data Mining, Pima Indians Diabetes Database, Ensemble Adaboost, Bagging

Abstract

Information and communication technology can be used by experts or doctors to interpret disease in a fast and accurate time. One application of information technology in the world of Health can be used to predict diabetes. This study will predict the pima indians diabetes database with the adaboost and bagging ensemble to produce higher accuracy in detecting diabetes. Data mining is a technology that can be used to assist companies in finding information that can be used from the data they have. The use of data mining is implemented for prediction, to predict what will happen in the future. In this study, comparing the use of adaboost and bagging ensembles against the nave Bayes classification algorithm, svm, and decision tree to produce the best accuracy and precision values from the Pima Indians Diabetes Database dataset. This research has known that the best accuracy value is using the SVM algorithm with the incorporation of ensemble bagging even though the change in the accuracy value does not increase significantly by 77.47%. However, the precision test resulted in the use of nave Bayes better without using an ensemble, either adaboost or bagging, with a value of 80.23%.

Keywords: *Data Mining, Pima Indians Diabetes Database, Ensemble Adaboost, Bagging*

1. Pendahuluan

Diabetes merupakan penyakit yang menjadi penyebab kebutaan, gagal ginjal dan serangan jantung hingga kematian. Menurut Organisasi International Diabetes Federation (IDF) pada tahun 2019 terdapat 463 juta orang di dunia yang menderita diabetes. Jumlah tersebut diprediksi meningkat di tahun 2030 sebesar 578 juta dan bahkan akan terus bertambah hinggan 700 juta di tahun 2045. Indonesia sendiri termasuk kedalam 10 negara dengan jumlah penderita diabetes tertinggi di dunia pada tahun 2019 (Kementrian kesehatan republik indonesia, 2020).

Penentuan jenis penyakit diabetes membutuhkan kemampuan ahli atau pakar mengenai (Riadi, 2017). Banyak orang yang terlambat mengetahui penyakit yang sedang dialami sehingga saat diperiksa sudah dalam kedaan yang parah (Hadi et al., 2018).

Meyikapi keadaan tersebut, deteksi penyakit diabetes sejak dini dibutuhkan sebagai salah satu upaya untuk mencegah kondisi yang lebih kompleks. Penggunaan teknologi dapat digunakan untuk meminimalisir kesalahan penaksiran. Dalam dunia medis, teknologi dapat digunakan oleh para pakar ataupun dokter untuk menafsirkan tentang penyakit dalam waktu yang cepat dan akurat (Gunawan et al., 2020).

Teknologi yang dapat digunakan untuk membantu dalam mendapatkan informasi dari data yang ada disebut sebagai data mining (Budiyasari et al., 2017). Data mining dapat digunakan untuk memprediksi kejadian di masa depan (Prajarini et al., 2016). Salah satunya memprediksi penyakit diabetes. Data mining memiliki teknis klasifikasi yang merupakan metode analisis untuk memprediksi label atau kelas dari data sampel dengan tujuan menemukan model model yang tepat sehingga dapat mengklasifikasi data baru yang akan dianalisis (Purwanto & Darmadi, 2018). Teknik AdaBoost dan Bagging diusulkan untuk meningkatkan kinerja klasifikasi pada algoritma data mining (Sugara & Subekti, 2019).

Algoritma Naïve Bayes, SVM, dan Decision Tree merupakan algoritma klasifikasi data mining. Algoritma Naïve Bayes digunakan untuk memprediksi peluang dimasa yang akan datang dengan mengidentifikasi peluang tertinggi (Retno Utari & Wibowo, 2020). Algoritma Support Vector Machine merupakan algoritma untuk mengenali pola dan menganalisis data dengan tujuan untuk memprediksi suatu sistem (Nurajijah et al., 2019). Algoritman Decision Tree adalah algoritma yang digunakan untuk menentukan atribut atribut mana dari suatu pohon keputusan akan dibagi (Algoritma et al., 2020).

Implementasi data mining untuk prediksi penyakit diabetes penelitian ini menghasilkan *decision tree* untuk mencegah penyakit diabetes sedini mungkin. Penelitian ini menghasilkan diagnose apakah orang terkena penyalit diabetes atau tidak (Putri et al., 2021).

Penelitian prediksi penyakit diabetes dengan mengoptimasi algoritma naïve bayes menggunakan teknik PSO dan stratified, hasil dari penelitian ini adalah Naïve Bayes mendapatkan akurasi 75,40% dan ROC 0,829%.

Penelitian dengan judul Perbandingan Kinerja Rule ZeroR Dan Function SMO Dengan T-Test untuk mendiagnosa penyakit diabetes militus. Hasil uji dengan Cross Validation mendapatkan akurasi 77,3% (Wiyono, 2016).

Prediksi penyakit diabetes menggunakan *fiture selection* naïve bayes dan korelasi pearson, Penggunaan algoritma korelasi pearson dibutuhkan untuk meningkatkan performa dari algoritma naïve bayes yang menghasilkan nilai akurasi dari 68,2% menjadi 79,13% pada informasi penyakit diabetes (Hanif & Khoirudin, 2020).

Penelitian sebelumnya belum mengadopsi teknik Ensemble AdaBoost dan Bagging yang mempunyai kegunaan untuk meningkatkan kinerja klasifikasi pada algoritma data mining.

Berdasarkan permasalahan yang telah dipaparkan, penelitian ini akan memprediksi pima Indians Diabetes Database dengan Ensemble AdaBoost Dan Bagging untuk menghasilkan akurasi yang lebih tinggi dalam mendeteksi penyakit diabetes.

2. Metode Penelitian

Metode pada penelitian ini menggunakan pendekatan kuantitatif, dimana dengan memberikan penekanan pengukuran dari data yang ada. Dalam melakukan penelitian, dilakukan tahap tahap penelitian seperti pada gambar 1 sebagai acuan. Data yang digunakan menggunakan *public dataset* yang berasal dari Kaggle dengan nama dataset *Pima Indians Diabetes Database* dengan halaman https://www.kaggle.com/uciml/pima-indians-diabetes-database (Learning, n.d.)

Gambar 1. Kerangka Penelitian

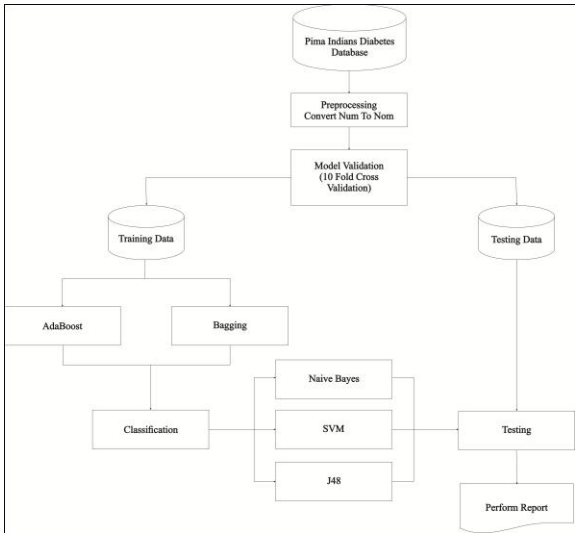
Pada *Pima Indians Diabetes Database* berkaitan dengan hasil diagnose diabetes yang dilakukan kepada pasien Wanita berusia minimal 21 tahun dengan 8 attribut predictor dan target outcome (0 or 1) seperti pada tabel 1. Dataset pima Indians diabetes database terdiri dari 768 data yang terbagi menjadi 8 attribut dan 2 kelas dengan jumlah kelas 1 (268) dan kelas 0 (500).

Tabel 1.Attribut dan outcome

No	Attribute Name	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3	BloodPressure	Diastolic blood pressure (mm Hg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2-Hour serum insulin (mu U/ml)
6	BMI	Body mass index (weight in kg/(height in m)^2)
7	DiabetesPedigreeFunction	Diabetes pedigree function
8	Age	Age (Years)
9	Outcome	Class Variable (0 or 1)

Sumber : Kaggle Pima Indians Diabetes Database (Learning, n.d.)

Setelah data didapatkan, kemudian dilakukan proses preprocessing dengan mengubah tipe data dari num ke nom. Data yang telah di konversi tipe datanya, kemdian masuk ke proses eksperimen model. Eksperimen model dilakukan dengan melakukan 3 klasifikasi percobaan, yaitu percobaan tanpa ensemble adaboost dan bagging, dengan adaboost, dan dengan bagging terhadap algoritma klasifikasi naïve bayes, svm, dan decision tree.



Sumber: Hasil Olahan Peneliti

Gambar 2. Eksperimen Model

Hasil dari eksperimen model akan dievaluasi dengan membandingkan antar hasil ujicoba dengan melihat hasil *confusion matriks.* Hasil dari *confusion matriks* akan dilihat nilai akurasi dan presisi yang dihasilkan. Nilai akurasi dan presisi akan dijadikan evaluasi model dengan mengambil nilai terbaiknya.

3. Hasil dan Pembahasan

3.1. Pengumpulan Data

Data pada penelitian ini menggunakan *dataset public* yang berasal dari Kaggle (https://www.kaggle.com/uciml/pima-indians-diabetes-database) dengan nama *dataset* yang digunakan menggunakan *public dataset* yang berasal dari Kaggle dengan nama dataset *Pima Indians Diabetes Database*. Dataset pima Indians diabetes database terdiri dari 768 data yang terbagi menjadi 8 attribut dan 2 kelas dengan jumlah kelas 1 (268) dan kelas 0 (500). Sebelum ke tahap eksperimen model, dilakukan proses perubahan tipedata pada *dataset* dari *Num* ke *Nom* menggunakan aplikasi WEKA 3.8.4.

3.2. Eksperimen Model

Eksperiman model menggunakan aplikasi WEKA 3.8.4. setelah dilakukan proses perubahan tipedata pada *dataset*, kemudian dilakukan proses eksperimen model dengan menggunakan *cross validation.* Hasil *cross validation* kemudian dihitung untuk melihat nilai akurasi dan presisi tiap pengujian dengan *confusion matrix*.

Tabel 1. *Confusion Matrix*

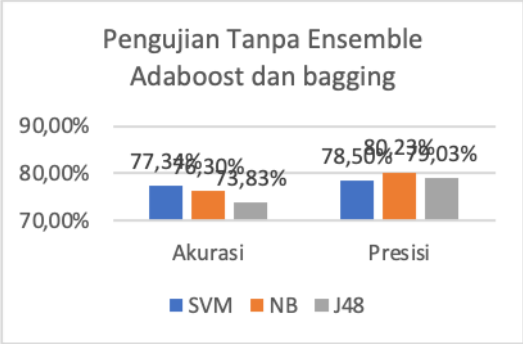
		Class Prediksi	
		Yes	No
Class Aktual	Yes	TP	FN
	No	FP	TN

$$Akurasi = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

Pengujian eksperimen dilakukan sebanyak 9 kali yang terbagi menjadi 3 kategori pengujian dengan membandingkan hasil kinerjan ensemble adaboost dan bagging terhadap algoritma klasifikasi naïve bayes, svm, dan J48. Pengujian kategori petama

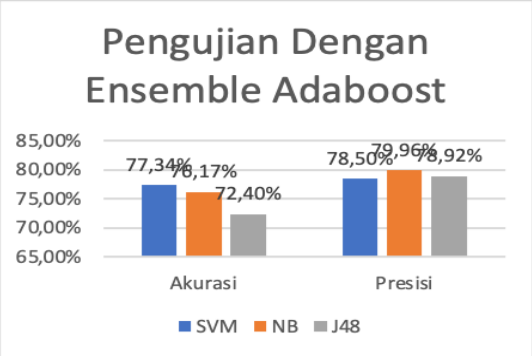
dilakukan sebanyak 3 kali dengan tanpa menggunakan ensemble adaboost dan bagging pada gambar 3. Dapat diketahui bahwa pengujian tanpa ensemble adaboost dan bagging, akurasi terbaik dengan menggunakan SVM sebesar 77,34%, namun presisi terbaik pada Naïve Bayes sebesar 80,23%.



Sumber: Hasil Olahan Peneliti

Gambar 3. Pengujian Tanpa Ensemble Adaboost dan bagging

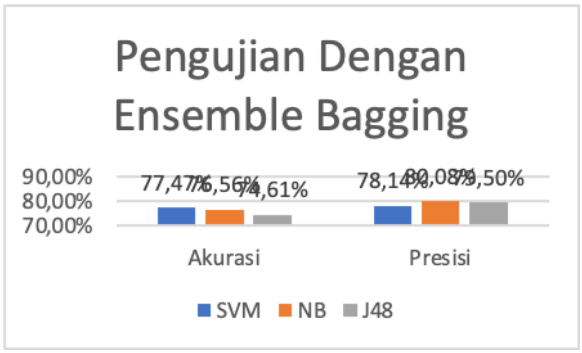
Pada pengujian kedua dilakukan pengujian *dataset* dengan algoritma klasifikasi yang digabungkan dengan ensemble adaboost. Hasil pengujian menunjukkan bahwa algoritma klasifikasi SVM yang digabungkan dengan ensemble adaboost memiliki akurasi 77,34%. Sedangkan jika melihat dari sisi presisi yang dihasilkan, dengan algoritma naïve bayes menghasilkan presisi tertinggi dengan nilai 79,96%.



Sumber: Hasil Olahan Peneliti

Gambar 4. Pengujian Dengan Ensemble Adaboost

Pada pengujian kedua dilakukan pengujian *dataset* dengan algoritma klasifikasi yang digabungkan dengan ensemble bagging. Hasil pengujian menunjukkan bahwa algoritma klasifikasi SVM yang digabungkan dengan ensemble adaboost memiliki akurasi 77,47%. Sedangkan jika melihat dari sisi presisi yang dihasilkan, dengan algoritma naïve bayes menghasilkan presisi tertinggi dengan nilai 80,08%.

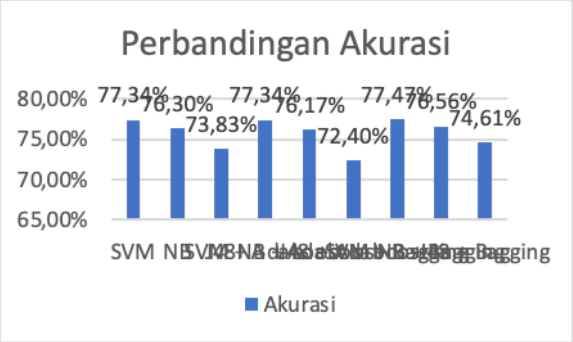


Sumber: Hasil Olahan Peneliti

Gambar 5. Pengujian Dengan Ensemble Bagging

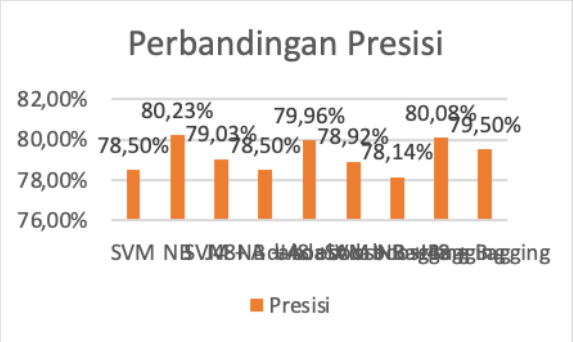
3.3. Evaluasi hasil

Dari ketiga kategori pengujian yang telah dilakukan, dapat diketahui bahwa nilai akurasi terbaik menggunakan algoritma SVM dengan penggabungan ensemble bagging meski perubahan nilai akurasinya tidak mengalami kenaikan yang signifikan sebesar 77,47% seperti pada gambar 6. Namun, pada pengujian presisi dihasilkan penggunaan naïve bayes lebih baik tanpa menggunakan ensemble baik adaboost maupun bagging dengan nilai 80,23% seperti gambar 7.



Sumber: Hasil Olahan Peneliti

Gambar 6. Perbandingan Akurasi



Sumber: Hasil Olahan Peneliti

Gambar 7. Perbandingan presisi

4. Kesimpulan

Dalam penelitian ini membandingkan penggunaan ensemble adaboost dan bagging terhadap algoritma klasifikasi naïve bayes, svm, dan decision tree untuk menghasilkan nilai akurasi dan presisi terbaik dari *dataset Pima Indians Diabetes Database*. Penelitian ini telah dapat diketahui bahwa nilai akurasi terbaik menggunakan algoritma SVM dengan penggabungan ensemble bagging meski perubahan nilai akurasinya tidak mengalami kenaikan yang signifikan sebesar 77,47%. Namun, pada pengujian presisi dihasilkan penggunaan naïve bayes lebih baik tanpa menggunakan ensemble baik adaboost maupun bagging dengan nilai 80,23%. Keterbatasan dalam penelitan adalah penelitian tentang klasifikasi penyakit diabetes yang menggunakan teknik optimasi metode belum banyak sehingga masih kurangnya referensi. Sedangkan evaluasi yang dapat dilakukan pada penelitian yang akan datang dapat dilakukan dengan menambahkan proses preprocessing agar dapat menghasilkan nilai akurasi dan presisi yang lebih baik. Penambahan teknik Particle Swarm Optimization untuk mengoptimalkan hasil klasifikasi serta dapat *development* program aplikasi.

Referensi

[1] Algoritma, P., Bayes, N., & Tree, D. A. N. D. (2020). *Perbandingan algoritma naïve bayes , svm, dan decision tree untuk klasifikasi sms spam*. *05*(02), 167–174. http://jurnal.univbinainsan.ac.id/index.php/jusim/article/download/956/631

[2] Budiyasari, V. N., Studi, P., Informatika, T., Teknik, F., Nusantara, U., & Kediri, P. (2017). Implementasi Data Mining Pada Penjualan kacamata Dengan Menggunakan Algoritma Apriori. *Indonesian Journal on Computer and Information Technology*, *2*(2), 31–39.

[3] Gunawan, M. I., Sugiarto, D., & Mardianto, I. (2020). Peningkatan Kinerja Akurasi Prediksi Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *6*(3), 280–284. https://jurnal.untan.ac.id/index.php/jepin/article/view/40718/75676587680

[4] Hadi, A. F., Setiawidayat, S., & Qustoniah, A. (2018). Perancangan Dan Pembuatan Aplikasi Sistem Pakar Untuk Mendiagnosa Penyakit Diabetes Mellitus Berbasis Android. *Jurnal WIDYA TEKNIKA*, *26*(1), 1–11.

https://publishing-widyagama.ac.id/ejournal-v2/index.php/widyateknika/article/viewFile/844/757

[5] Hanif, M. B., & Khoirudin. (2020). Sistem Aplikasi Prediksi Penyakit Diabetes Menggunakan Fiture. *Pengembangan Rekayasa Dan Teknologi*, *16*(2), 199–205.

[6] Kementrian kesehatan republik indonesia. (2020). Tetap Produktif, Cegah Dan Atasi Diabetes Mellitus. In *pusat data dan informasi kementrian kesehatan RI*.

[7] Learning, U. M. (n.d.). *Pima Indians Diabetes Database*. Kaggle. https://www.kaggle.com/uciml/pima-indians-diabetes-database

[8] Nurajijah, N., Ningtyas, D. A., & Wahyudi, M. (2019). Klasifikasi Siswa Smk Berpotensi Putus Sekolah Menggunakan Algoritma Decision Tree, Support Vector Machine Dan Naive Bayes. *Jurnal Khatulistiwa Informatika*, *7*(2), 85–90. https://doi.org/10.31294/jki.v7i2.6839

[9] Prajarini, D., Tinggi, S., Rupa, S., Desain, D., & Indonesia, V. (2016). Perbandingan Algoritma Klasifikasi Data Mining Model C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Diabetes. *Informatics Journal*, *1*(3), 137. http://ejournal.nusamandiri.ac.id/index.php/techno/article/view/217/193

[10] Purwanto, A., & Darmadi, E. A. (2018). Perbandingan Minat Siswa Smu Pada Metode Klasifikasi Menggunakan 5 Algoritma. *IKRAITH-INFORMATIKA*, *2*(1), 43–47. http://journals.upi-yai.ac.id/index.php/ikraith-informatika/article/download/158/79

[11] Putri, sanni ucha, Irawan, E., & Rizky, F. (2021). Implementasi Data Mining Untuk Prediksi Penyakit Diabetes. *KESATRIA(Jurnal Penerapan Sistem Informasi Dan Manajemen*, *2*(1), 39–46. https://tunasbangsa.ac.id/pkm/index.php/kesatria/article/viewFile/56/56

[12] Retno Utari, D., & Wibowo, A. (2020). Pemodelan Prediksi Status Keberlanjutan Polis Asuransi Kendaraan dengan Teknik Pemilihan Mayoritas Menggunakan Algoritma-Algoritma Klasifikasi Data Mining. *Prosiding Seminar Nasional Teknoka*, *5*(2502), 19–24. https://doi.org/10.22236/teknoka.v5i.391

[13] Riadi, A. (2017). Penerapan Metode Certainty Factor Untuk Sistem Pakar Diagnosa Penyakit Diabetes Melitus Pada Rsud Bumi Panua Kabupaten Pohuwato. *ILKOM Jurnal Ilmiah*, *9*(3), 309–316. https://doi.org/10.33096/ilkom.v9i3.162.309-316

[14] Sugara, B., & Subekti, A. (2019). Penerapan Support Vector Machine (SVM) Pada Small Dataset Untuk Deteksi Dini Gangguan Autisme. *Jurnal Pilar Nusa Mandiri*, *15*(2), 177–182. https://doi.org/10.33480/pilar.v15i2.649

[15] Wiyono, S. (2016). Perbandingan Kinerja Rule ZeroR dan Function SMO dengan T-Test dalam Pengklasifikasian Diagnosis Penyakit Diabetes Mellitus. *Emitor: Jurnal Teknik Elektro*, *16*(1), 23–25. https://doi.org/10.23917/emitor.v16i1.2679