STUDI KOMPARASI ALGORITMA ID3 DAN ALGORITMA NAIVE BAYES UNTUK KLASIFIKASI PENYAKIT DIABETES MELLITUS

Nunu Nurdiana [1], Abijar Algifari [2]
[1,2]Fakultas Teknik, Universitas Majalengka
Email: [1]abijaralgifari69@gmail.com, [2]nun@unma.ac.id

ABSTRAK

Penyakit diabetes mellitus salah satu penyakit yang mematikan, merupakan penyakit gangguan metabolik menahun akibat pankreas tidak memproduksi cukup insulin atau tubuh tidak dapat menggunakan insulin yang diproduksi secara efektif. Berdasarkan data history penderita diabetes dapat dibuat rekomendasi prediksi penyakit diabetes yang dapat membantu tenaga kesehatan. Klasifikasi merupakan salah satu teknik dari data mining yang dapat digunakan untuk membantu prediksi hasil klasifikasi penyakit diabetes. Klasifikasi dilakukan menggunakan Algoritma ID3 dan Algoritma Naive Bayes dengan bahasa pemrograman python menggunakan aplikasi web open source yaitu Jupyter Notebook. Penelitian ini bertujuan membuat klasifikasi dan menerapkan klasifikasi data mining. Hasil klasifikasi data di evaluasi dengan menggunakan Confusion Matrix dan kurva ROC untuk mengetahui tingkat hasil akurasi menggunakan algoritma ID3 yaitu sebesar 74% dan nilai AUC dari kurva ROC adalah 0.788 sedangkan Algoritma Naive Bayes sebesar 76% nilai AUC dari kurva ROC 0.794 sehingga dapat dikatakan bahwa Algoritma Naive Bayes memiliki hasil prediksi yang baik dalam memprediksi penyakit diabetes meliitus.

Kata kunci: Diabetes Mellitus, Klasifikasi, ID3, Naive Bayes.

1. PENDAHULUAN

1.1. Latar Belakang

Diabetes merupakan penyakit gangguan metabolik menahun akibat pankreas tidak memproduksi cukup insulin atau tubuh tidak dapat menggunakan insulin yang diproduksi secara efektif. Berdasarkan sebab yang mendasari kemunculannya, diabetes menjadi beberapa golangan atau tipe. Di antara tipe-tipe diabetes yan**g** termasuk tipe utama adalah diabetes tipe-1 dan tipe-2. Diabetes tipe-1 biasanya mengenai anak- anak dan remaja. Diabetes Mellitus juga merupakan salah satu penyebab utama penyakit ginjal dan kebutaan pada usia di bawah 65 tahun, dan juga amputasi (Marshall dan Flyvbjerg, 2006). Selain itu, diabetes juga menjadi penyebab terjadinya amputasi (yang bukan disebabkan oleh trauma), disabilitas, hingga kematian. Dampak lain dari diabetes adalah mengurangi usia harapan hidup sebesar 5-10 tahun. Usia harapan hidup penderita DM tipe 2 yang mengidap penyakit mental serius, seperti Skizofrenia, bahkan 20% lebih rendah dibandingkan dengan populasi umum. (Garnita, 2012).

Data WHO menunjukkan bahwa angka kejadian penyakit tidak menular pada tahun 2004 yang mencapai 48,30% sedikit lebih besar dari angka kejadian penyakit menular, yaitu sebesar 47,50%. Bahkan penyakit tidak menular menjadi penyebab kematian nomor satu di dunia (63,50%). (Faktor Risiko Diabetes Mellitus di Indonesia (Analisis Data Sakerti 2007) (Garnita, 2012). WHO memperkirakan bahwa, secara global, 422 juta orang dewasa berusia di atas 18 tahun hidup dengan diabetes pada tahun 2014. Jumlah terbesar orang dengan diabetes diperkirakan berasal dari Asia Tenggara dan Pasifik Barat, terhitung sekitar setengah kasus diabetes di dunia. Di seluruh dunia, jumlah penderita diabetes telah meningkat secara substansial antara tahun 1980 dan 2014, meningkat dari 108 juta menjadi 422 juta atau sekitar empat kali lipat. Teknik klasifikasi secara manual sudah tidak lagi efektif digunakan karena jumlah data penderita diabetes mellitus yang banyak dan perlu dilakukan seleksi fitur-fitur pada dataset sehingga membutuhkan waktu yang cukup lama dan tingkat akurasi data yang baik. Diagnosis terhadap penyakit Diabetes Mellitus secara medis sendiri masih mengalami kesulitan dan bahkan mengalami reduksi data. Data medis yang memiliki sejumlah fitur yang tidak relevan, dan redundant dapat memberikan pengaruh terhadap kualitas dari diagnosis penyakit (Nurahman & Prihandoko 2019).Untuk mendukung mengenai diagnosis perlu menggunakan teknik klasifikasi data mining berbasis komputer agar dapat menggali informasi dari dataset informasi penyakit Diabetes Mellitus.

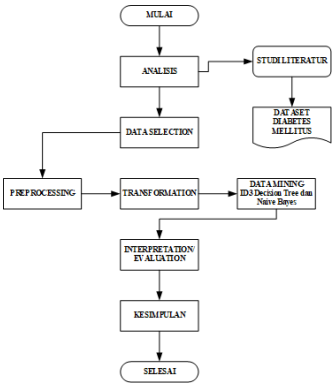
Klasifikasi adalah proses dari mencari suatu himpunan model (fungsi) yang dapat mendeskripsikan dan membedakan kelas-kelas data atau konsep-konsep, dengan tujuan dapat menggunakan model tersebut untuk memprediksi kelas dari suatu objek yang mana kelasnya belum diketahui (Rani, 2015). Untuk memulai sebuah klasifikasi dibutuhkan suatu metode data mining pada penelitian ini menggunakan algoritma ID3 dan Naive Bayes. Algoritma ID3 merupakan sebuah metode yang digunakan untuk membuat pohon keputusan. Algoritma pada metode ini menggunakan

konsep dari entropi informasi sedangkan algoritma Naive Bayes merupakan metode yang membagi permasalahan ke dalam sebuah kelas-kelas berdasarkan ciri-ciri persamaan dan perbedaan dengan menggunakan statistik yang bisa memprediksi probabilitas sebuah kelas.

Oleh karena itu, penelitian ini dilakukan untuk membantu menyelesaikan permasalahan tersebut dengan data mining untuk klasifikasi penyakit diabetes mellitus, diperlukan suatu metode atau teknik yang dapat mengolah data-data yang sudah ada. Salah satu metodenya menggunakan teknik data mining. Penggunaan data mining dengan algoritma ID3 dan Naive Bayes sebagai pilihan untuk klasifikasi penyakit diabetes mellitus dapat menjadi alternatif pilihan yang tepat, tetapi sampai saat ini belum diketahui algoritma yang paling akurat klasifikasi penyakit diabetes. Pada penelitian ini akan dilakukan komparasi data mining algoritma ID3 dan Naive Bayes untuk mengetahui algoritma yang memiliki akurasi yang lebih tinggi dalam klasifikasi penyakit diabetes mellitus. Bedasarkan beberapa hal yang dijelaskan diatas maka untuk penelitian Tugas Akhir ini peneliti akan memberikan judul "STUDI KOMPARASI ALGORITMA ID3 DAN ALGORITMA NAIVE BAYES UNTUK KLASIFIKASI PENYAKIT DIABETES MELLITUS"

1.2. Metodologi Penelitian

Metodologi penelitian yang digunakan dalam Studi Komparasi Algoritma ID3 Dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus sebagai berikut:



Gambar 1 Kerangka Penelitian

Tahapan kerangka penelitian ini yaitu:

a. Analisis

Dalam proses analisis ini, hal yang perlu diperhatikan yaitu mencari data terlebih dahulu melalui observasi, wawancara, dan studi literatur. Sehingga data tersebut akan mudah ditemukan untuk dijadikan bahan analisis ke dalam perhitungan Algoritma ID3 Dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus.

b. Data Selection

Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakuan sebelum tahap penggalian informasi dalam KDD dimulai data hasil seleksi yang akan digunakan untu proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional. Pada Pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (discovery) akan dilakukan. Hasil seleksi disimpan dalam suatu berkas, terpisah dari basis data operasional.

c. Preprocessing

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan pross cleaning pada data yang menjadi fokus KDD. proses cleaning mencakup antara lain membuang dupliasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). juga dilakukan proses ercihment, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data ata informasi eksternal.

d. Transformasi.

Merupakan proses integrasi pada data yang telah dipilih, sehingga data sesuai untuk proses data mining. Merupakan proses yang sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data. Data transformasi tersebut akan dipilih dalam perhitungan data seleksi.

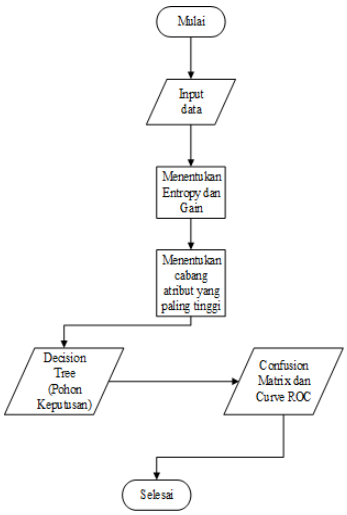
e. Data Mining.

Pemilihan tugas data mining merupakan pemilihan goal dari proses KDD misalnya karakterisasi, klasifikasi, regresi, Clustering, asosiasi, dan lain-lain. Pemilihan tugas data mining merupakan pemilihan goal dari proses KDD misalnya karakterisasi, klasifikasi, regresi, Clustering, asosiasi, dan lain-lain. Pemilihan teknik, metode atau Algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

f. Evalution.

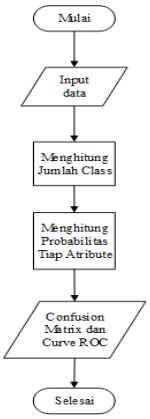
Yaitu penerjemahan pola-pola yang dihasilkan dari data mining. Pola informasi yang dihasilkan perlu ditampilkan dalam bentuk yang mudah dimengerti. Tahap ini melakukan pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

1. Flowchart Algoritma ID3



Gambar 2 Flowchart Algoritma ID3

2. Flowchart Algoritma Naive Bayes



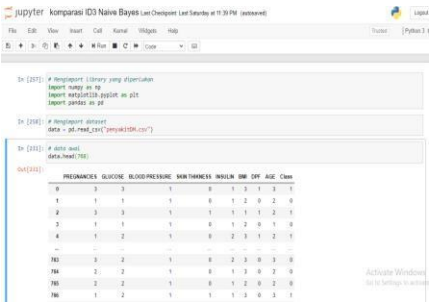
Gambar 3 Flowchart Algoritma Naive Bayes

2. PEMBAHASAN

Di sini peneliti menggunakan Anaconda sebagai tools karena di dalam anaconda sudah terdapat Jupyter Notebook. Jupyter Notebook biasa juga disebut pengembangan dari Ipython atau Interactive Python. Jupyter Notebook ini suatu editor dalam bentuk web aplikasi yang berjalan di localhost komputer, adapun beberapa hal yang dapat dilakukan oleh Jupyter Notebook seperti menulis kode python, equations, visualisasi dan bisa juga sebagai markdown editor.

1. Input Data

Sebelum data diolah kedalam data mining, Input data selection hasil diskritisati atribut import kedalam Jupyter Notebook. Berikut adalah tampilan Import data menggunakan bahasa pemrograman Python pada Jupyter Notebook.

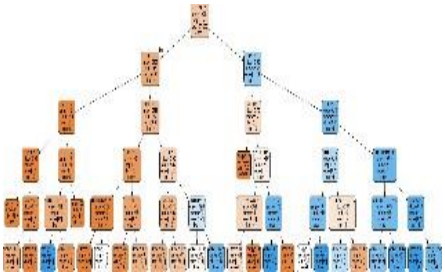


Gambar 4 Input Data

2. Visualize Decison Tree

Mengunakan fungsi export_graphviz Scikit-learn untuk menampilkan pohon dalam notebook Jupyter. Dan untuk plotting tree, import graphviz dan pydotplus.

a. Ouput Decision Tree Tampilan output decision tree pada jupyter notebook.



Gambar 5 Output Decision Tree

Pada Gambar diatas terdapat 28 rule merupakan hasil dari klasifikasi dengan menggunakan model algoritma ID3 dengan Decision Tree dapat dijelaskan sebagai berikut:

1. R1

IF Glucose <= 2.50 AND Bmi <= 1.50 AND Age <= 2.50 AND Dpf <= 0.50 THEN Class: 0

2. R2

IF Glucose <= 2.50 AND Bmi <= 1.50 AND Age <= 2.50 AND Dpf > 0.50 Age <= 1.50 THEN

Class: 0

3. R3

IF Glucose <= 2.50 AND Bmi <= 1.50 AND Age <= 2.50 AND Dpf > 0.50 Age > 1.50 THEN

Class: 0

4. R4

IF Glucose <= 2.50 AND Bmi <= 1.50 AND Age > 2.50 AND Dpf <= 0.50 AND Pregnancies <= 2.50 THEN Class: 1

5. R5

IF Glucose <= 2.50 AND Bmi <= 1.50 AND Age > 2.50 AND Dpf <= 0.50 AND Pregnancies >

2.50 THEN Class: 0

6. R6

IF Glucose <= 2.50 AND Bmi <= 1.50 AND Age > 2.50 AND Dpf <= 0.50 THEN Class: 0

7. R7

IF Glucose <= 2.50 AND Bmi > 1.50 AND Age <= 1.50 AND Bmi <= 2.50 AND Pregnancies <= 2.50 THEN Class: 0

8. R8

IF Glucose <= 2.50 AND Bmi > 1.50 AND Age <= 1.50 AND Bmi <= 2.50 AND Pregnancies >

2.50 THEN Class: 0

9. R9

IF Glucose <= 2.50 AND Bmi > 1.50 AND Age <= 1.50 AND Bmi <= 2.50 AND Pregnancies > 2.50 THEN Class: 0

10. R10

IF Glucose <= 2.50 AND Bmi > 1.50 AND Age <= 1.50 AND Bmi > 2.50 AND Glucose <= 1.50 THEN Class: 0

11. R11

IF Glucose <= 2.50 AND Bmi > 1.50 AND Age <= 1.50 AND Bmi > 2.50 AND Glucose > 1.50 THEN Class: 0

12. R12

IF Glucose <= 2.50 AND Bmi > 1.50 AND Age > 1.50 AND Dpf <= 0.50 AND Age <= 2.50 THEN Class: 0

13. R13

IF Glucose <= 2.50 AND Bmi > 1.50 AND Age > 1.50 AND Dpf <= 0.50 AND Age > 2.50 THEN Class: 0

14. R14

IF Glucose <= 2.50 AND Bmi > 1.50 AND Age > 1.50 AND Dpf > 0.50 AND Insulin <= 1.50 THEN Class: 0

R15

IF Glucose <= 2.50 AND Bmi >1.50 AND Age > 1.50 AND Dpf > 0.50 AND Insulin > 1.50 THEN Class: 1

16. R16

IF Glucose > 2.50 AND Bmi <= 2.50 AND Age <= 1.50 THEN Class: 0

17. R17

IF Glucose > 2.50 AND Bmi <= 2.50 AND Age <= 1.50 AND Insulin <= 1.50 AND Blood Pressure <= 2.50 THEN Class: 0

18. R18

IF Glucose > 2.50 AND Bmi <= 2.50 AND Age <= 1.50 AND Insulin <= 1.50 AND Blood Pressure > 2.50 THEN Class: 0

19. R19

IF Glucose > 2.50 AND Bmi <= 2.50 AND Age <= 1.50 AND Insulin > 1.50 AND Dpf <= 0.50 THEN Class: 1

20. R20

IF Glucose > 2.50 AND Bmi <= 2.50 AND Age <= 1.50 AND Insulin > 1.50AND Dpf > 0.50 THEN Class: 0

21. R21

IF Glucose > 2.50 AND Bmi > 2.50 AND Age <= 1.50 AND Insulin <= 2.50 AND Dpf <= 0.50 THEN Class: 0

22. R22

IF Glucose > 2.50 AND Bmi > 2.50 AND Age <= 1.50 AND Insulin <= 2.50 AND Dpf > 0.50 THEN Class: 1

23. R23

IF Glucose > 2.50 AND Bmi > 2.50 AND Age <= 1.50 AND Insulin > 2.50 AND Blood Pressure <= 1.50 THEN Class: 1

24. R24

IF Glucose > 2.50 AND Bmi > 2.50 AND Age <= 1.50 AND Insulin > 2.50 AND Blood Pressure > 1.50 THEN Class: 0

25. R25

IF Glucose > 2.50 AND Bmi > 2.50 AND Age > 1.50 AND Skin Thikness <= 0.50 Blood Pressure <= 1.50 THEN Class: 1

26. R26

IF Glucose > 2.50 AND Bmi > 2.50 AND Age > 1.50 AND Skin Thikness <= 0.50 Blood Pressure > 1.50 THEN Class: 1

27. R27

IF Glucose > 2.50 AND Bmi > 2.50 AND Age > 1.50 AND Skin Thikness > 0.50 Age <= 2.50 THEN Class: 1

28. R28

IF Glucose > 2.50 AND Bmi > 2.50 AND Age > 1.50 AND Skin Thikness > 0.50 Age > 2.50 THEN Class: 1

3. Confusion Matrix

Mengunakan fungsi classification_report dan confusion_matrix pada Library Scikit-learn metrics untuk menampilkan precision, recall f1-score, suport dan accuracy.

a. Confusion Matrix Decision Tree

```
In [244]: # confusion matrix
          predictions = dt.predict(x_test)
          from sklearn.metrics import classification_report,confusic
          print(confusion_matrix(y_test,predictions))
          print(classification_report(y_test,predictions))

          [[118  12]
           [ 37  25]]
                        precision    recall  f1-score   support

                     0       0.76      0.91      0.83       130
                     1       0.68      0.40      0.51        62

              accuracy                           0.74       192
             macro avg       0.72      0.66      0.67       192
          weighted avg       0.73      0.74      0.72       192
```

Gambar 6 ID3 Decision Tree

Dari hasil classification_report dan confusion_matrix diatas diketahui terdapat 118 TP (True Prositif), 12 FN (False Negatif), 37 FP (False Positif), dan 25 (True Negatif).

b. Confusion Matrix Naive Bayes

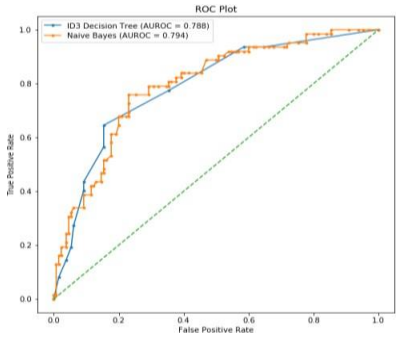
```
In [249]: from sklearn.metrics import classification_report, confusion_matrix
          print(confusion_matrix(y_test,y_pred))
          print(classification_report(y_test,y_pred))

          [[107  23]
           [ 24  38]]
                        precision    recall  f1-score   support

                     0       0.82      0.82      0.82       130
                     1       0.62      0.61      0.62        62

              accuracy                           0.76       192
             macro avg       0.72      0.72      0.72       192
          weighted avg       0.75      0.76      0.75       192
```

Gambar 7 Naive Bayes

Dari hasil classification_report dan confusion_matrix diatas diketahui terdapat 107 TP (True Prositif), 23 FN (False Negatif), 24 FP (False Positif), dan 38 (True Negatif).

4. Hasil Pengujian Curve AUROC

Mengunakan fungsi roc_curve dan roc_auc_score pada Library Scikit-learn metrics untuk menampilkan Nilai AUC dan Curve ROC.



Gambar 8 Curve AUROC

Berdasarkan data pada Curva diatas, diketahui bahwa nilai akurasi Algoritma ID3 adalah 74% dengan nilai AUC 0.788, sedangkan nilai akurasi Naive Bayes 76% dan nilai AUC 0.794 sedangkan hasil pengujian dari prediksi diabetes mellitus hasilnya termasuk Fair Classification.

3. KESIMPULAN

Berdasarkan hasil evaluasi dan implementasi yang sudah dilakukan, maka kesimpulan dari Penelitian Tugas Akhir dengan judul "Studi Komparasi Algoritma ID3 Dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus" yaitu sebagai berikut:

a) Evaluasi Kinerja Algoritma ID3 dan Algoritma Naive Bayes dalam melakukan klasifikasi penyakit diabetes mellitus dengan menganalisa tingkat akurasinya menggunakan Confusion Matrix dan Kurva ROC. Berdasarkan hasil pengukuraan tingkat akurasi kedua algoritma tersebut, diketahui bahwa nilai akurasi Algoritma ID3 adalah 74% dengan nilai AUC 0.788, sedangkan nilai akurasi Naive Bayes 76% dan nilai AUC 0.794.

b) Komparasi Algoritma ID3 dan Algoritma Naive Bayes pada klasifikasi data mining penyakit diabetes mellitus dapat disimpulkan bahwa dengan menggunakan model Naive Bayes lebih tinggi tingkat akurasinya, dengan peningkatan akurasi sebesar 2% dan peningkatan nilai AUC sebesar 0.006 sedangkan hasil pengujian dari prediksi diabetes mellitus hasilnya termasuk Fair Classification.

PUSTAKA

Dita Garnita,Faktor Resiko DM di Indonesia. Universitas Indonesia.2007

Metisen, Benri Melpa dan Herlina Latipa Sari. Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokkan Penjualan Produk Pada Swalayan Fadhila. ISSN: 1858-2680. Jakarta: Jurnal Media Infotama Vol. 11, No. 2, September 2015: 110-118. Diambil dari: jurnal.unived.ac.id/index.php/jmi/article/download/258/237/. (18 Mei 2016)

Misnadiarly. 2006. Diabetes Melitus Gangren, Ulcer, Infeksi, Mengenali gejala, Menanggulangi, dan Mencegah komplikasi. Jakarta: Pustaka Obor Populer.

PERKENI., 2011. Konsensus Pengelolaan dan Pencegahan Diabetes MelitusTipe 2 di Indonesia. Jakarta.

Playing Game Dengan Metode Finite State Machine. Universitas Malikussaleh, 1-11.

Prajarini, D., 2016. Perbandingan Algoritma Klasifikasi Data Mining Untuk Prediksi Penyakit Kulit. Informatics Journal Vol.1 No.3 (2016)

Pramudiono, I. 2007. Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data. http://www.ilmukomputer.org/wp-content/uploads/2006/08/ikodata mining.zip. 26 April 2017 (19:54)

Prasetyo, E. 2012. Data Mining Konsep dan Aplikasi Menggunakan MATLAB. Andi. Yogyakarta.

Purnamasari, D., 2009. Diagnosis dan Klasifikasi Diabetes Melitus. Di Dalam :Buku Ajar Ilmu Penyakit Dalam. Jilid 3 Edisi V. Jakarta: Pusat Penerbit Departemen Ilmu Penyakit Dalam FK UI, hal. 1880-1883.

WHO. 2016. Diabetes. World Health Organization. (online)diakses 23 Mei 2016. http://www.who.int/mediacentre/factsheets/fs312/en/-46