

**PEMILIHAN ATRIBUT TERBAIK
PADA PREDIKSI PENYAKIT DIABETES
BERDASAR ALGORITMA KLASIFIKASI ID3**



SKRIPSI

**Disusun Sebagai Salah Satu Syarat
untuk Memperoleh Gelar Sarjana Komputer
pada Departemen Ilmu Komputer/ Informatika**

Disusun Oleh :

Muhamad Subhan Efendi

24010313130081

**DEPARTEMEN ILMU KOMPUTER/ INFORMATIKA
FAKULTAS SAINS DAN MATEMATIKA
UNIVERSITAS DIPONEGORO**

2018

HALAMAN PERNYATAAN KEASLIAN SKRIPSI

Saya yang bertanda tangan di bawah ini :

Nama : Muhamad Subhan Efendi

NIM : 24010313130081

Judul : Pemilihan Atribut Terbaik Pada Prediksi Penyakit Diabetes Berdasar Algoritma
Klasifikasi ID3

Dengan ini saya menyatakan bahwa dalam tugas akhir/ skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar kesarjanaan di suatu Perguruan Tinggi dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan di dalam daftar pustaka...

Semarang, 28 Maret 2018
METERAI
TEMPEL
TGL. 28/03/2018
19AF4AEF985E3462
6000
ENAM RIBU RUPIAH
Muhamad Subhan Efendi
24010313130081

HALAMAN PENGESAHAN

Judul : Pemilihan Atribut Terbaik Pada Prediksi Penyakit Diabetes Berdasar Algoritma
Klasifikasi ID3

Nama : Muhamad Subhan Efendi

NIM : 24010313130081

Telah diujikan pada sidang tugas akhir tanggal 14 Maret 2018 dan dinyatakan lulus pada
14 Maret 2018



Drs. Parolita Kasumaningrum, S.Si, M.Kom
NIP. 198104202005012001

Semarang, 28 Maret 2018

Panitia Penguji Tugas Akhir
Ketua,

Drs. Eko Adi Sarwoko, M.Kom
NIP. 196511071992031003

HALAMAN PENGESAHAN

Judul : Pemilihan Atribut Terbaik Pada Prediksi Penyakit Diabetes Berdasar Algoritma
Klasifikasi ID3

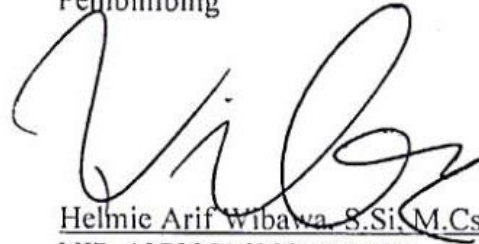
Nama : Muhamad Subhan Efendi

NIM : 24010313130081

Telah diujikan pada sidang tugas akhir dan dinyatakan lulus pada 14 Maret 2018

Semarang, 28 Maret 2018

Pembimbing



Helmic Arif Wibawa, S.Si, M.Cs
NIP. 197805162003121001

ABSTRAK

Penyakit diabetes atau sering disebut dengan penyakit kencing manis adalah suatu penyakit gangguan metabolik menahun yang ditandai oleh kadar glukosa dalam darah yang melebihi nilai normal. Penyakit diabetes sering disebut sebagai *silent killer* dengan mengacu pada banyaknya yang tidak menyadari bahwa dirinya terkena penyakit diabetes sampai diketahui sudah kronis. Hal ini memicu peningkatan jumlah penderita diabetes dari tahun ke tahun. Penelitian ini mencoba menerapkan pemilihan atribut terbaik dalam memprediksi penyakit diabetes berdasar algoritma klasifikasi *Data Mining*. Untuk pemilihan atribut terbaik digunakan algoritma seleksi atribut *Correlation based Feature Selection (CFS)* dan *Information Gain* Sedangkan algoritma klasifikasi yang digunakan adalah algoritma ID3. Berdasarkan hasil penelitian ini diperoleh bahwa performa tertinggi dicapai ketika algoritma ID3 menggunakan 5 atribut yaitu glukosa darah puasa, glukosa darah 2 jam, glukosa urin puasa, glukosa urin 2 jam, dan aseton urin puasa. Dimana kelima atribut tersebut diperoleh menggunakan algoritma *Correlation based Feature Selection (CFS)* dengan nilai rata-rata akurasi sebesar 84.77, nilai rata-rata *sensitifity* sebesar 87.18, nilai rata-rata *specificity* sebesar 82.37, dan nilai rata-rata FNR sebesar 12.82.

Kata Kunci : Penyakit Diabetes, Data Mining, ID3, Seleksi Atribut

ABSTRACT

Diabetes is a chronic metabolic disease disorder characterized by levels of glucose in the blood that exceeds normal value. Diabetes is often called as a silent killer with reference to many who do not realize that he was exposed to diabetes until it is said to be chronic. This cause an increase of number of diabetics from year to year. This research tried to apply the best attribute selection in predicting diabetes based on Data Mining classification algorithm. For the best attributes selection used Correlation based Feature Selection (CFS) and Information Gain attribute selection algorithm while the classification algorithm used is ID3 algorithm. Based on results of this research, it is found that the highest performance is obtained when the ID3 algorithm uses 5 attributes namely fasting blood glucose, blood glucose 2 hours, fasting urine glucose, urine glucose 2 hours, fasting urine acetone. That attributes are obtained using Correlation based Feature Selection (CFS) algorithm with an average accuracy is 84.77, average sensitivity is 87.18, average of specificity is 82.37, and average of FNR is 12.82.

Keyword : Diabetes. Data Mining, ID3, Attribute Selection

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas berkat dan kasihNya sehingga penulis dapat menyelesaikan laporan tugas akhir yang berjudul “Pemilihan Atribut Terbaik Pada Prediksi Penyakit Diabetes Berdasar Algoritma Klasifikasi ID3” dengan baik dan lancar. Laporan tugas akhir ini disusun sebagai salah satu syarat untuk memperoleh gelar sarjana strata satu pada Departemen Ilmu Komputer/ Informatika Fakultas Sains dan Matematika Universitas Diponegoro Semarang. Dalam penyusunan tugas akhir ini penulis banyak mendapat bimbingan, bantuan, dan dukungan dari berbagai pihak. Oleh karena itu, dengan segala kerendahan hati, penulis menyampaikan terimakasih kepada:

1. Prof. Dr. Widowati, M.Si, selaku Dekan FSM UNDIP
2. Dr. Retno Kusumaningrum, S.Si, M.Kom, selaku Ketua Departemen Ilmu Komputer/ Informatika
3. Helmie Arif Wibawa, S.Si, M.Cs, selaku Koordinator Tugas Akhir dan Dosen Pembimbing

Penulis menyadari bahwa dalam laporan ini masih banyak kekurangan baik dari penyampaian materi maupun isi dari materi itu sendiri. Hal ini dikarenakan keterbatasan kemampuan dan pengetahuan dari penulis. Oleh karena itu, kritik dan saran yang bersifat membangun sangat penulis harapkan. Semoga laporan tugas akhir ini dapat bermanfaat bagi penulis dan juga pembaca pada umumnya.

Semarang, 14 Maret 2018

Muhamad Subhan Efendi

DAFTAR ISI

| | |
|--|------|
| HALAMAN JUDUL | i |
| HALAMAN PERNYATAAN KEASLIAN SKRIPSI..... | ii |
| HALAMAN PENGESAHAN | iii |
| HALAMAN PENGESAHAN | iv |
| ABSTRAK | v |
| ABSTRACT | vi |
| KATA PENGANTAR..... | vii |
| DAFTAR ISI | viii |
| DAFTAR GAMBAR..... | xi |
| DAFTAR TABEL | xiii |
| DAFTAR LAMPIRAN | xv |
| BAB I PENDAHULUAN | 1 |
| 1.1. Latar Belakang..... | 1 |
| 1.2. Rumusan Masalah | 2 |
| 1.3. Tujuan dan Manfaat..... | 3 |
| 1.4. Ruang Lingkup | 3 |
| BAB II STUDI PUSTAKA | 5 |
| 2.1. Tinjauan Pustaka | 5 |
| 2.2. Penyakit Diabetes | 6 |
| 2.3. <i>Data Mining</i> | 7 |
| 2.4. <i>Imbalance Data</i> | 8 |
| 2.5. Seleksi Atribut | 10 |
| 2.5.1. <i>Correlation based Feature Selection (CFS)</i> | 11 |
| 2.5.2. <i>Information Gain</i> | 12 |
| 2.6. <i>Decision Tree Iterative Dichotomiser 3 (ID3)</i> | 13 |
| 2.7. <i>Cross Validation</i> | 15 |
| 2.8. <i>Confusion Matrix</i> | 15 |
| 2.9. <i>Model Waterfall</i> | 17 |
| 2.10. Pemodelan Fungsional..... | 19 |
| 2.11. Pemodelan Data..... | 20 |
| BAB III METODOLOGI PENELITIAN | 21 |

| | | |
|----------------------------------|---|----|
| 3.1 | Data..... | 21 |
| 4.2. | <i>Data Preprocessing</i> | 23 |
| 4.2.1. | <i>Data cleaning</i> | 23 |
| 4.2.2. | <i>Data integration</i> | 24 |
| 4.2.3. | <i>Data Transformation</i> | 24 |
| 4.2.4. | <i>Data Selection</i> | 27 |
| 4.2.4.1. | Penanganan <i>Imbalance Data</i> | 27 |
| 4.2.4.2. | Seleksi Atribut..... | 30 |
| 4.3. | Pembagian Data Latih dan Data Uji | 34 |
| 4.4. | Proses <i>Data Mining</i> menggunakan Algoritma ID3 | 34 |
| 4.5. | Analisa Kebutuhan Aplikasi..... | 37 |
| 4.5.1. | Kebutuhan Fungsional dan Non-Fungsional..... | 37 |
| 4.5.2. | Pemodelan Data | 37 |
| 4.5.3. | Pemodelan Fungsional | 39 |
| 4.5.3.1. | DFD Level 0..... | 39 |
| 4.5.3.2. | DFD Level 1..... | 39 |
| 4.6. | Desain Aplikasi | 41 |
| 4.6.1. | Deskripsi Aplikasi..... | 41 |
| 4.6.2. | Desain Antarmuka | 41 |
| 4.6.3. | Desain Fungsi | 45 |
| BAB IV HASIL DAN PEMBAHASAN..... | | 50 |
| 4.1. | Implementasi Aplikasi..... | 50 |
| 4.1.1. | Lingkungan Implementasi | 50 |
| 4.1.2. | Implementasi Antarmuka..... | 50 |
| 4.1.3. | Implementasi Fungsi..... | 53 |
| 4.2. | Pengujian Fungsional Aplikasi..... | 54 |
| 4.3. | Skenario Pengujian | 54 |
| 4.3.1. | Skenario 1 | 55 |
| 4.3.2. | Skenario 2 | 55 |
| 4.3.3. | Skenario 3 | 55 |
| 4.3.4. | Skenario 4 | 55 |
| 4.3.5. | Skenario 5 | 56 |
| 4.3.6. | Skenario 6 | 56 |

| | | |
|----------------------------------|-------------------------------------|----|
| 4.4. | Pembahasan Skenario Pengujian | 56 |
| 4.4.1. | Pembahasan Skenario 1 | 56 |
| 4.4.2. | Pembahasan Skenario 2 | 57 |
| 4.4.3. | Pembahasan Skenario 3 | 59 |
| 4.4.4. | Pembahasan Skenario 4 | 61 |
| 4.4.5. | Pembahasan Skenario 5 | 62 |
| 4.4.6. | Pembahasan Skenario 6 | 64 |
| 4.5. | Analisa Hasil Pengujian | 66 |
| BAB V KESIMPULAN DAN SARAN | | 73 |
| 5.1. | Kesimpulan..... | 73 |
| 5.2. | Saran | 73 |
| DAFTAR PUSTAKA..... | | 74 |
| LAMPIRAN-LAMPIRAN | | 76 |

DAFTAR GAMBAR

| | |
|--|----|
| Gambar 2.1 Diagram Proses <i>Data Mining</i> | 8 |
| Gambar 2.2 Ilustrasi <i>K-Fold Cross Validation</i> dengan $k=3$ | 15 |
| Gambar 2.3 Model <i>Waterfall</i> | 18 |
| Gambar 3.1 Garis Besar Permasalahan Penelitian | 21 |
| Gambar 3.2 Ilustrasi <i>7-fold cross validation</i> | 34 |
| Gambar 3.3 Ilustrasi Pembentukan <i>Node</i> | 36 |
| Gambar 3.4 <i>Entity Relationship Diagram</i> Aplikasi Prediksi Diabetes | 38 |
| Gambar 3.5 DFD Level 0 | 39 |
| Gambar 3.6 DFD Level 1 | 40 |
| Gambar 3.7 Desain Halaman Awal | 42 |
| Gambar 3.8 Desain Halaman <i>Import Data</i> | 42 |
| Gambar 3.9 Desain Halaman Seleksi Atribut | 43 |
| Gambar 3.10 Desain Halaman Pelatihan (<i>Data Mining</i>) dan Pengujian | 44 |
| Gambar 3.11 Desain Halaman Prediksi | 45 |
| Gambar 3.12 Desain Fungsi <i>Cleaning</i> | 45 |
| Gambar 3.13 Desain Fungsi <i>Transform</i> | 46 |
| Gambar 3.14 Desain Fungsi Seleksi Atribut | 46 |
| Gambar 3.15 Desain Fungsi <i>Undersampling</i> | 47 |
| Gambar 3.16 Desain Fungsi <i>K-Fold</i> | 48 |
| Gambar 3.17 Desain Fungsi Pelatihan (<i>Data Mining</i>) | 48 |
| Gambar 3.18 Desain Fungsi Pengujian | 49 |
| Gambar 3.19 Desain Fungsi Prediksi | 49 |
| Gambar 4.1 Implementasi Halaman Awal | 51 |
| Gambar 4.2 Implementasi Halaman <i>Import Data</i> | 51 |
| Gambar 4.3 Implementasi Halaman Seleksi Atribut | 52 |
| Gambar 4.4 Implementasi Halaman Pelatihan (<i>Data Mining</i>) dan Pengujian | 53 |
| Gambar 4.5 Implementasi Halaman Prediksi | 53 |
| Gambar 4.6 Diagram Skenario Pengujian | 54 |
| Gambar 4.7 Grafik Hasil Pengujian Skenario 1 | 57 |
| Gambar 4.8 Grafik Hasil Pengujian Skenario 2 | 59 |
| Gambar 4.9 Grafik Hasil Pengujian Skenario 3 | 61 |

| | |
|--|----|
| Gambar 4.10 Grafik Hasil Pengujian Skenario 4 | 62 |
| Gambar 4.11 Grafik Hasil Pengujian Skenario 5 | 64 |
| Gambar 4.12 Grafik Hasil Pengujian Skenario 6 | 66 |
| Gambar 4.13 Grafik perbandingan hasil pengujian pada penggunaan data tidak imbang dan data imbang | 67 |
| Gambar 4.14 Grafik perbandingan akurasi penggunaan CFS dan IG pada data tidak imbang | 68 |
| Gambar 4.15 Grafik perbandingan <i>sensitifity</i> penggunaan CFS dan IG pada data tidak Imbang | 68 |
| Gambar 4.16 Grafik perbandingan <i>specificity</i> penggunaan CFS dan IG pada data tidak Imbang | 68 |
| Gambar 4.17 Grafik perbandingan FNR penggunaan CFS dan IG pada data tidak imbang ... | 69 |
| Gambar 4.18 Grafik perbandingan akurasi penggunaan CFS dan IG pada data imbang ... | 70 |
| Gambar 4.19 Grafik perbandingan <i>sensitifity</i> penggunaan CFS dan IG pada data imbang | 70 |
| Gambar 4.20 Grafik perbandingan <i>specificity</i> penggunaan CFS dan IG pada data imbang ... | 71 |
| Gambar 4.21 Grafik perbandingan FNR penggunaan CFS dan IG pada data imbang | 71 |

DAFTAR TABEL

| | |
|--|----|
| Tabel 2.1 Daftar Referensi | 5 |
| Tabel 2.2 Tabel Informasi Atribut | 7 |
| Tabel 2.3 Tabel <i>Confusion Matrix</i> 2 Kelas | 16 |
| Tabel 2.4 Contoh Hasil Pengujian | 17 |
| Tabel 2.5 Contoh Hasil <i>Confusion Matrix</i> | 17 |
| Tabel 2.6 Simbol-simbol DFD | 19 |
| Tabel 2.7 Simbol-simbol ERD | 20 |
| Tabel 3.1 Keterangan Atribut | 22 |
| Tabel 3.2 Contoh Data Diabetes (diambil 10 data pertama) | 22 |
| Tabel 3.3 Contoh Data dengan <i>Missing Value</i> | 23 |
| Tabel 3.4 Contoh Data dengan <i>Missing Value</i> Sudah Terisi | 24 |
| Tabel 3.5 Kriteria Pengendalian Diabetes Dalam mg/dL | 25 |
| Tabel 3.6 Tabel Pembagian Kriteria Atribut | 26 |
| Tabel 3.7 Data Sudah Melalui Proses Diskritisasi | 27 |
| Tabel 3.8 Data Sudah Melalui Proses <i>Mapping</i> | 27 |
| Tabel 3.9 Komposisi Jumlah Data Setelah Proses <i>Undersampling</i> | 32 |
| Tabel 3.10 Pembobotan Kombinasi Atribut Terbaik Dengan CFS | 32 |
| Tabel 3.11 Pembobotan Atribut Dengan <i>Information Gain</i> | 33 |
| Tabel 3.12 Contoh Komposisi Jumlah untuk Kategori Umur | 35 |
| Tabel 3.13 Kebutuhan Fungsional Aplikasi | 37 |
| Tabel 3.14 Keterangan Entitas | 38 |
| Tabel 4.1 Implementasi Fungsi | 54 |
| Tabel 4.2 Tabel Hasil Pengujian Skenario1 | 56 |
| Tabel 4.3 Hasil Pemilihan Atribut Menggunakan CFS pada Skenario 2 | 57 |
| Tabel 4.4 Tabel Hasil Pengujian Skenario 2 untuk penggunaan 1 s.d 3 atribut | 58 |
| Tabel 4.5 Tabel Hasil Pengujian Skenario 2 untuk penggunaan 4 s.d 6 atribut | 58 |
| Tabel 4.6 Tabel Hasil Pengujian Skenario 2 untuk penggunaan 7 s.d 9 atribut | 58 |
| Tabel 4.7 Tabel Hasil Pengujian Skenario 2 untuk penggunaan 10 s.d 11 atribut | 58 |
| Tabel 4.8 Hasil Pemilihan Atribut dengan <i>Information Gain</i> pada Skenario 3 | 59 |
| Tabel 4.9 Tabel Hasil Pengujian Skenario 3 untuk penggunaan 1 s.d 3 atribut | 60 |
| Tabel 4.10 Tabel Hasil Pengujian Skenario 3 untuk penggunaan 4 s.d 6 atribut | 60 |

| | |
|--|----|
| Tabel 4.11 Tabel Hasil Pengujian Skenario 3 untuk penggunaan 7 s.d 9 atribut | 60 |
| Tabel 4.12 Tabel Hasil Pengujian Skenario 3 untuk penggunaan 10 s.d 11 atribut | 60 |
| Tabel 4.13 Tabel Hasil Pengujian Skenario 4 | 61 |
| Tabel 4.14 Hasil Pemilihan Atribut dengan CFS pada Skenario 5 | 62 |
| Tabel 4.15 Tabel Hasil Pengujian Skenario 5 untuk penggunaan 1 s.d 3 atribut | 63 |
| Tabel 4.16 Tabel Hasil Pengujian Skenario 5 untuk penggunaan 4 s.d 6 atribut | 63 |
| Tabel 4.17 Tabel Hasil Pengujian Skenario 5 untuk penggunaan 7 s.d 9 atribut | 63 |
| Tabel 4.18 Tabel Hasil Pengujian Skenario 5 untuk penggunaan 10 s.d 11 atribut | 63 |
| Tabel 4.19 Hasil Pemilihan Atribut dengan <i>Information Gain</i> pada Skenario 6 | 64 |
| Tabel 4.20 Tabel Hasil Pengujian Skenario 6 untuk penggunaan 1 s.d 3 atribut | 65 |
| Tabel 4.21 Tabel Hasil Pengujian Skenario 6 untuk penggunaan 4 s.d 6 atribut | 65 |
| Tabel 4.22 Tabel Hasil Pengujian Skenario 6 untuk penggunaan 7 s.d 9 atribut | 65 |
| Tabel 4.23 Tabel Hasil Pengujian Skenario 6 untuk penggunaan 10 s.d 11 atribut | 65 |
| Tabel 4.24 Perbandingan hasil pengujian pada penggunaan data tidakimbang dan data imbang | 66 |
| Tabel 4.25 Perbandingan rata-rata hasil pengujian performa ketika seleksi atribut CFS dan <i>Information Gain</i> pada data tidakimbang | 67 |
| Tabel 4.26 Perbandingan hasil pengujian ketika menggunakan seleksi atribut CFS dan <i>Information Gain</i> pada dataimbang | 70 |

DAFTAR LAMPIRAN

| | |
|---|----|
| Lampiran 1. Tabel Pengujian <i>Blackbox</i> | 77 |
|---|----|

BAB I

PENDAHULUAN

Bab ini dijelaskan mengenai latar belakang, rumusan masalah, tujuan dan manfaat, serta ruang lingkup dalam pembuatan tugas akhir ini.

1.1. Latar Belakang

International Diabetes Federation (IDF) pada tahun 2013 membuat estimasi bahwa jumlah pengidap diabetes di dunia mencapai 382 juta orang. Diperkirakan dari 382 juta orang tersebut, sekitar 175 juta di antaranya belum terdiagnosa, sehingga terancam berkembang tanpa disadari dan tanpa pencegahan. Jumlah tersebut diperkirakan akan naik menjadi 592 juta orang pada tahun 2035 (Kemenkes RI, 2014).

Di Indonesia sendiri jumlah penderita diabetes cukup tinggi, yaitu sekitar 12 juta orang pada tahun 2013. Jumlah tersebut ternyata meningkat daripada tahun-tahun sebelumnya. Pada tahun 2007-2013, Risesdas (Riset Kesehatan Dasar) melakukan survei untuk menghitung proporsi penderita diabetes untuk usia 15 tahun ke atas. Survei diambil dari data orang yang pernah didiagnosa menderita penyakit diabetes oleh dokter dan yang belum pernah didiagnosa oleh dokter tetapi dalam 1 bulan terakhir mengalami gejala-gejala awal diabetes. Hasil survei tersebut mendapatkan jumlah penderita diabetes pada tahun 2013 meningkat dua kali lipat dibandingkan tahun 2007 (Kemenkes RI, 2014).

Peningkatan jumlah penderita diabetes dikarenakan diabetes dikenal sebagai *silent killer*. Hal ini mengacu pada banyaknya yang tidak menyadari bahwa dirinya terkena penyakit diabetes. Penderita biasanya diketahui terjangkit penyakit ini ketika sudah terjadi komplikasi tanpa adanya penanganan di awal (Kemenkes RI, 2014). Untuk menekan jumlah penderita penyakit diabetes yang semakin bertambah, bisa dilakukan deteksi dini yang dapat dilakukan oleh tenaga ahli.

Untuk melakukan deteksi dini penyakit diabetes, dapat dikembangkan suatu sistem untuk memprediksi penyakit dengan memanfaatkan berbagai metode. Salah satu metode yang dapat digunakan yaitu metode *data mining* dengan prinsip klasifikasi. Metode ini dapat mengolah data dalam jumlah besar yang nantinya digunakan untuk mendapatkan hasil prediksi. Seperti yang telah diterapkan dalam penelitian sebelumnya dengan menggunakan algoritma C4.5 (Jasri, 2017).

Salah satu metode yang dapat digunakan dalam klasifikasi data mining adalah *decision tree* (pohon keputusan). Metode ini telah diterapkan untuk memprediksi tingkat kelulusan mahasiswa (Kamagi dan Hansun, 2014), untuk memprediksi loyalitas pelanggan (Santoso, 2013), serta di bidang medis pernah diterapkan untuk memprediksi penyakit kanker payudara (Mutmainah, 2015).

Pada penelitian sebelumnya digunakan *data mining* yaitu algoritma *Naive Bayes* untuk memprediksi ketepatan kelulusan mahasiswa (Rozzaqi, 2015) dan algoritma ID3 (*Iterative Dichotomiser 3*) untuk memprediksi penyakit diabetes (Sathya dan Rajesh, 2016). Akurasi dari algoritma ID3 untuk memprediksi diabetes berada pada kisaran 63. Namun untuk algoritma *Naive Bayes* yang digunakan untuk memprediksi ketepatan kelulusan mahasiswa mendapatkan nilai akurasi yang cukup tinggi setelah pemilihan atribut dengan *information gain* dengan 3 atribut, yaitu pada kisaran 89.79 yang pada awalnya menggunakan 13 atribut didapat akurasi sebesar 83.07. Disini dapat dilihat bahwa pemilihan atribut yang digunakan dapat mempengaruhi hasil prediksi. Pada kasus lain, dilakukan perbandingan stabilitas penggunaan beberapa algoritma seleksi atribut pada beberapa algoritma *classifier* dengan hasilnya didapatkan algoritma *Correlation based Feature Selection* (CFS) merupakan algoritma yang paling stabil dan mendapatkan nilai akurasi yang lebih tinggi (Djatna dan Morimoto, 2011). Dengan nilai akurasi dari algoritma ID3 yang hanya mencapai 63 pada penelitian sebelumnya (Sathya dan Rajesh, 2016), dapat diterapkan algoritma *Correlation based Feature Selection* (CFS) dan *Information Gain* dalam pemilihan atribut untuk meningkatkan performa algoritma ID3.

Berdasar penelitian tersebut maka dapat dibuat suatu penerapan pemilihan atribut terbaik menggunakan seleksi atribut *Correlation based Feature Selection* (CFS) atau *Information Gain* pada prediksi penyakit diabetes berdasarkan algoritma ID3.

1.2. Rumusan Masalah

Berdasar latar belakang yang telah dijelaskan, maka rumusan masalah untuk penelitian ini adalah :

1. Bagaimana menerapkan seleksi atribut *Information Gain* atau *Correlation based Feature Selection* pada prediksi penyakit diabetes berdasar algoritma klasifikasi ID3.

2. Metode seleksi atribut mana yang menghasilkan performa lebih bagus di antara *Information Gain* dan *Correlation based Feature Selection* untuk kasus ini.
3. Bagaimana performa algoritma ID3 setelah ditambahkan algoritma seleksi atribut *Information Gain* dan *Correlation based Feature Selection*

1.3. Tujuan dan Manfaat

Tujuan dari pembuatan Tugas Akhir ini adalah :

1. Menerapkan seleksi atribut *Information Gain* atau *Correlation based Feature Selection* dalam prediksi penyakit diabetes berdasar algoritma klasifikasi ID3.
2. Untuk mencari performa terbaik algoritma ID3 setelah penerapan algoritma seleksi atribut *Information Gain* dan *Correlation based Feature Selection*

Manfaat dari pembuatan Tugas Akhir ini adalah :

1. Penelitian ini dapat menjadi referensi dalam pemilihan atribut yang dapat digunakan dalam penelitian selanjutnya mengenai diabetes
2. Penelitian ini dapat menjadi referensi untuk penelitian *data mining* selanjutnya.

1.4. Ruang Lingkup

Ruang lingkup dari penelitian ini adalah sebagai berikut :

1. Penelitian ini difokuskan hanya untuk memprediksi penyakit diabetes saja tanpa adanya tindakan lanjutan untuk menanganinya
2. Penelitian ini melakukan penyeimbangan data menggunakan *cluster based undersampling*
3. Penelitian ini menggunakan algoritma seleksi atribut *Correlation based Feature Selection* (CFS) dan *Information Gain* dalam pemilihan atribut terbaik.
4. Data yang digunakan pada penelitian ini berasal dari Rumah Sakit Pusat Pertamina Jakarta yang memuat data dari tahun 2013 sampai tahun 2015

1.5. Sistematika Penulisan

Sistematika penulisan yang digunakan dalam tugas akhir ini terbagi dalam beberapa pokok bahasan, yaitu :

BAB I PENDAHULUAN

Bab ini membahas latar belakang masalah, rumusan masalah, tujuan dan manfaat, ruang lingkup, serta sistematika penulisan dalam penyusunan tugas akhir.

BAB II STUDI PUSTAKA

Bab ini menyajikan tinjauan pustaka dan landasan teori yang digunakan dalam penyusunan tugas akhir

BAB III METODOLOGI PENELITIAN

Bab ini menyajikan tahapan-tahapan dalam penyusunan tugas akhir serta perancangan aplikasi dalam tugas akhir ini.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan implementasi aplikasi serta pembahasan analisis dan evaluasi hasil pengujian

BAB V PENUTUP

Bab ini berisi kesimpulan dari uraian yang telah dijabarkan pada bab-bab sebelumnya dan saran untuk pengembangan penelitian lebih lanjut.