PERBANDINGAN TEKNIK *UNDERSAMPLING* DAN *OVERSAMPLING* PADA KLASIFIKASI DATA PASIEN DIABETES MELLITUS (DM) DENGAN MENGGUNAKAN ALGORITMA *NAIVE BAYES CLASSIFIER* (NBC)

SKRIPSI

Diajukan untuk Memenuhi Salah Satu Syarat Mencapai Gelar Strata Satu

Program Studi Informatika

Disusun Oleh :

DEWI PRASETYAN DRAJATI

M0510018

PROGRAM STUDI INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SEBELAS MARET

SURAKARTA

2016

SKRIPSI

PERBANDINGAN TEKNIK *UNDERSAMPLING* DAN *OVERSAMPLING* PADA KLASIFIKASI DATA PASIEN DIABETES MELLITUS (DM) DENGAN MENGGUNAKAN ALGORITMA *NAIVE BAYES CLASSIFIER* (NBC)

Disusun Oleh :

DEWI PRASETYAN DRAJATI

M0510018

ditulis dan diajukan untuk memenuhi sebagian persyaratan memperoleh gelar Strata Satu Program Studi Informatika

PROGRAM STUDI INFORMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS SEBELAS MARET

SURAKARTA

2016

SKRIPSI

PERBANDINGAN TEKNIK UNDERSAMPLING DAN OVERSAMPLING PADA KLASIFIKASI DATA PASIEN DIABETES MELLITUS (DM) DENGAN MENGGUNAKAN ALGORITMA KLASIFIKASI NAIVE BAYES CLASSIFIER (NBC)

Disusun Oleh:

DEWI PRASETYAN DRAJATI

M0510018

Telah disetujui oleh pembimbing pada tanggal

Pembimbing I	Pembimbing II
Ristu Saptono, S.Si.,M.T.	**Winarno, S.Si., M.Eng**
NIP. 19790210 200212 1 001	NIP. 19820520 200604 1 001

SKRIPSI

PERBANDINGAN TEKNIK *UNDERSAMPLING* DAN *OVERSAMPLING* PADA KLASIFIKASI DATA PASIEN DIABETES MELLITUS (DM) DENGAN MENGGUNAKAN ALGORITMA *NAIVE BAYES CLASSIFIER* (NBC)

Disusun oleh:

DEWI PRASETYAN DRAJATI

M0510018

Telah dipertahankan di depan Dewan Penguji

pada tanggal:

Susunan Dewan Penguji:

1. **Ristu Saptono, S.Si.,M.T.** ( )
   NIP. 19790210 20021 21 001
2. **Winarno, S.Si., M.Eng** ( )
   NIP. 19820520 200604 1 001
3. **Esti Suryani, S.Si.,M.Kom.** ( )
   NIP. 19761129 200812 2 001
4. **Abdul Aziz, S.Kom., M.Cs.** ( )
   NIP. 19810413 200501 1 001

Disahkan oleh

Kepala Program Studi Informatika

Drs. Bambang Harjito, M.App.Sc.,Ph.D.

NIP. 19621130 199103 1 002

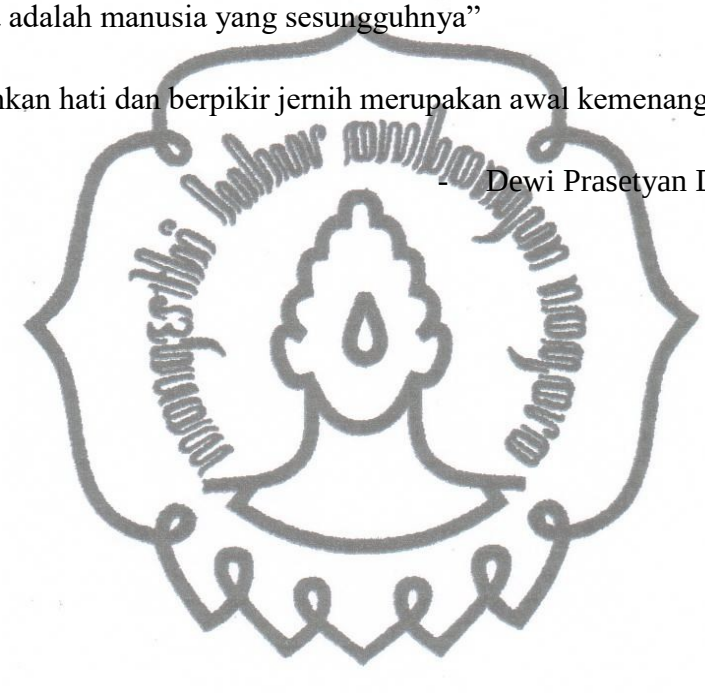
HALAMAN MOTTO

"Perjalanan dan penantian yang begitu panjang akan memiliki akhir ketika kita mulai melangkah"

"Berani menghadapi hidup dengan segala resiko di belakangnya, merupakan tanda bahwa kita adalah manusia yang sesungguhnya"

"Merendahkan hati dan berpikir jernih merupakan awal kemenangan dalam hidup"

- Dewi Prasetyan Drajati -



HALAMAN PERSEMBAHAN

"*Skripsi ini secara khusus saya persembahkan kepada eyang kakung dan eyang putri dan kepada ayah ibu saya yang senantiasa memberikan kekuatan dan dukungan kepada saya untuk menyelesaikan skripsi ini*."

KATA PENGANTAR

Segala puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa, karena oleh karunia-Nya penulis dapat menyelesaikan penulisan skripsi sebagai salah satu syarat untuk mendapatkan gelar strata satu Informatika Universitas Sebelas Maret Surakarta.

Penulis juga mengucapkan terimakasih kepada beberapa pihak yang telah mendukung dan membantu penulis untuk menyelesaikan penyusunan skripsi ini. Penulis mengucapkan terima kasih kepada:

1. Orang tua penulis, yang selalu menguatkan dan memberikan motivasi, doa, dan dukungan kepada penulis.
2. Bapak Drs. Bambang Harjito, M.App.Sc.,Ph.D. selaku Kepala Program Studi Informatika, Fakultas MIPA, Universitas Sebelas Maret, yang telah memberikan motivasi kepada penulis untuk menyelesaikan penyusunan skripsi ini.
3. Bapak Ristu Saptono, S.Si.,M.T. selaku dosen pembimbing I, yang senantiasa sabar dalam memberikan arahan dan juga dukungan kepada penulis selama proses penyusunan skripsi berlangsung.
4. Bapak Winarno, S.Si., M.Eng selaku dosen pembimbing II yang telah memberikan arahan dan masukkan kepada penulis dalam menyelesaikan penyusunan skripsi.
5. Bapak- Ibu Dosen Program Studi Informatika FMIPA UNS atas ilmu yang sudah diberikan kepada penulis selama menempuh masa studi di Program Studi Informatika.
6. Teman- teman Program Studi Informatika yang selalu memberikan semangat kepada penulis.

Surakarta,

Penulis

ABSTRACT

Imbalance dataset is a big problem inside a classification process. Most of the classification algorithms tend to classify the majority instances and ignore the minority ones. It can cause the misclassification of the minority instances and make the precision and recall of this minority data become low. In order to resolve this kind of problem there will be done both undersampling and oversampling process to make the dataset balance. In this proposed research there will be used undersampling and oversampling techniques to balance the number of majority and minority instances from Diabetic Mellitus (DM) patient data. The undersampling process will be done by using Spreadsubsample technique, meanwhile the oversampling process will be done by using Synthetic Minority Oversampling Technique (SMOTE). The other techniques used in this research are Backward Greedy Stepwise for atributte selection and Naive Bayes Classifier (NBC) algorithm for data classification. The results of the proposed research are the higher average of the accuration value for training data of DM patients that was proceed by using undersampling method than the accuration value for training data that was proceed by using oversampling method and the higher average of the accuration, precision, and also recall values for the testing data of DM patients that was proceed by using oversampling method.

Keywords : undersampling; oversampling; Spreadsubsample; SMOTE; Backward Greedy Stepwise; NBC

ABSTRAK

Ketidakseimbangan data merupakan masalah serius dalam suatu proses klasifikasi. Sebagian besar algoritma klasifikasi cenderung hanya melakukan klasifikasi pada data anggota kelas mayor sehingga mengabaikan proses klasifikasi pada data anggota kelas minor. Hal ini dapat berpengaruh terhadap perolehan nilai presisi dan *recall* pada data anggota kelas minoritas sehingga menjadi sangat kecil. Untuk menyelesaikan permasalahan ketidakseimbangan data yang terjadi, maka pada penelitian ini akan dilakukan baik proses *undersampling* maupun proses *oversampling* untuk menyeimbangkan data. Proses *undersampling* dilakukan menggunakan teknik *Spreadsubsample*, sedangkan proses *oversampling* dilakukan menggunakan *Synthetic Minority Oversampling Technique* (SMOTE). Data yang digunakan pada penelitian ini adalah data pasien Diabetes Mellitus (DM). Teknik lain yang juga digunakan dalam penelitian yang diusulkan ini adalah *Backward Greedy Stepwise* untuk melakukan pencarian terhadap atribut yang berpengaruh positif terhadap hasil klasifikasi dan algoritma *Naive Bayes Classifier* (NBC) untuk melakukan klasifikasi data. Hasil dari penelitian yang telah dilakukan adalah nilai rata- rata akurasi tertinggi yang diberikan oleh data latih dari data pasien DM yang diproses dengan metode *undersampling* lebih tinggi dibandingkan dengan akurasi data latih yang diproses dengan metode *oversampling*, meskipun nilai rata- rata tertinggi pada nilai presisi dan nilai *recall* yang diberikan oleh data latih yang diproses dengan metode *undersampling* lebih rendah. Nilai rata- rata tertinggi akurasi, presisi, dan *recall* dari data uji yang dikenai proses *oversampling* lebih tinggi dibandingkan dengan nilai rata- rata tertinggi data uji yang dikenai proses *undersampling*.

Keywords : *undersampling*; *oversampling*; *Spreadsubsample*; SMOTE; *Backward Greedy Stepwise*; NBC

DAFTAR ISI

SKRIPSI i
SKRIPSI i
SKRIPSI ii
SKRIPSI iii
HALAMAN MOTTO iv
HALAMAN PERSEMBAHAN v
KATA PENGANTAR vi
ABSTRACT vii
ABSTRAK viii
DAFTAR ISI ix
DAFTAR TABEL xi
DAFTAR GAMBAR xii
DAFTAR LAMPIRAN xiii
BAB I PENDAHULUAN 14
1.1 Latar Belakang 14
1.2 Rumusan Masalah 15
1.3 Batasan Masalah 16
1.4 Tujuan Penelitian 16
1.5 Manfaat Penelitian 16
1.6 Sistematika Penulisan 16
BAB II TINJAUAN PUSTAKA 18
2.1 Dasar Teori 18
2.1.1 *Undersampling* 18
2.1.2 *Oversampling* 19

2.1.3 *Confusion Matrix* 21
2.1.4 Akurasi, Presisi, dan *Recall* 22
2.1.5 Seleksi Fitur 22
2.1.6 *Naïve Bayes Classifier* (NBC) 23
2.2 Penelitian Terkait 24
BAB III METODOLOGI PENELITIAN 28
3.1 Tahap Pengumpulan Data 29
3.2 Pembagian Data Pasien DM ke Dalam Sepuluh Kelompok Sampel . 35
3.3 *Data Preprocessing* 35
3.3.1 Proses *Undersampling* dan *Oversampling* Data Pasien DM 35
3.3.2 Proses Seleksi Atribut 36
3.3.3 Proses Pembagian Data Latih dan Data Uji 36
3.4 Klasifikasi Data Pasien DM 37
3.5 Evaluasi Hasil 37
BAB IV HASIL DAN PEMBAHASAN 38
4.1 Pengumpulan Data 38
4.2 Pembagian Data Pasien DM ke Dalam Sepuluh Kelompok Sampel . 38
4.3 *Data Preprocessing* 39
4.3.1 *Sampling* 39
4.3.2 Seleksi Atribut 45
4.3.3 Pembagian Data Latih dan Data Uji 49
4.4 Klasifikasi Data Pasien DM 51
4.5 Evaluasi 51
BAB V KESIMPULAN DAN SARAN 58
5.1 Kesimpulan 58
5.2 Saran 59
DAFTAR PUSTAKA 60
LAMPIRAN 63

DAFTAR TABEL

Tabel 2.1. *Confusion Matrix* 22
Tabel 2.2. Penelitian Terkait 26
Tabel 3.1. Atribut Data Pasien DM..... 29
Tabel 3.2. Penelitian Terkait 29
Tabel 3.3. Persentase Data Latih dan Data Uji 36
Tabel 4.1. Perbandingan Jumlah Data Mayor Terhadap Jumlah Data Minor..... 40
Tabel 4.2. Contoh Atribut Sampel 1 41
Tabel 4.3. Hasil Perhitungan VDM 44
Tabel 4.4. Perbandingan Jumlah Data Minor Terhadap Jumlah Data Mayor..... 45
Tabel 4.5. Atribut Terpilih Data Pasien dengan Proses *Undersampling* 45
Tabel 4.6. Atribut Terpilih Data Pasien DM Dengan Proses *Oversampling* 46
Tabel 4.7. Hasil Klasifikasi Data Pasien DM Pada Sampel 1..... 48
Tabel 4.8. Data Latih..... 49
Tabel 4.9. Data Uji 50
Tabel 4.10. Rata- Rata Hasil Klasifikasi Data Latih (Metode *Undersampling*) ... 51
Tabel 4.11. Rata- Rata Hasil Klasifikasi Data Latih (Metode *Oversampling*) 51
Tabel 4.12. Rata- Rata Hasil Klasifikasi Data Uji (Metode *Undersampling*) 52
Tabel 4.13. Rata- Rata Hasil Klasifikasi Data Uji (Metode *Oversampling*)..... 52

DAFTAR GAMBAR

Gambar 2.1. Ilustrasi Proses *Undersampling* 18
Gambar 2.2. Ilustrasi Proses *Oversampling* 20
Gambar 3.1. Skema Metodologi Penelitian 28
Gambar 4.1. Pembagian Data Pasien ke Dalam 10 Kelompok Sampel 39
Gambar 4.2. Perbandingan Nilai Akurasi Data Latih 53
Gambar 4.3. Perbandingan Nilai Rata- Rata Presisi Data Latih 53
Gambar 4.4. Perbandingan Nilai Rata- Rata *Recall* Data Latih 54
Gambar 4.5. Perbandingan Nilai Rata- Rata Akurasi Data Uji 54
Gambar 4.6. Perbandingan Nilai Rata- Rata Presisi Data Uji 55
Gambar 4.7. Perbandingan Nilai Rata- Rata *Recall* Data Uji 55
Gambar 4.8. Perbandingan Nilai Rata- Rata *Recall* Data Uji 55

DAFTAR LAMPIRAN

Lampiran 1. Hasil Klasifikasi Data Latih 66% (Metode *Undersampling*) 63
Lampiran 2. Hasil Klasifikasi Data Latih 75% (Metode *Undersampling*) 64
Lampiran 3. Hasil Klasifikasi Data Latih 80% (Metode *Undersampling*) 64
Lampiran 4. Hasil Klasifikasi Data Latih 90% (Metode *Undersampling*) 64
Lampiran 5. Hasil Klasifikasi Data Latih 66% (Metode *Oversampling*)..... 64
Lampiran 6. Hasil Klasifikasi Data Latih 75% (Metode *Oversampling*)..... 64
Lampiran 7. Hasil Klasifikasi Data Latih 80% (Metode *Oversampling*)..... 64
Lampiran 8. Hasil Klasifikasi Data Latih 90% (Metode *Oversampling*)..... 64
Lampiran 9. Hasil Klasifikasi Data Uji 34% (Metode *Undersampling*)..... 64
Lampiran 10. Hasil Klasifikasi Data Uji 25% (Metode *Undersampling*)..... 64
Lampiran 11. Hasil Klasifikasi Data Uji 20% (Metode *Undersampling*)..... 64
Lampiran 12. Hasil Klasifikasi Data Uji 10% (Metode *Undersampling*)..... 64
Lampiran 13. Hasil Klasifikasi Data Uji 34% (Metode *Oversampling*)..... 64
Lampiran 14. Hasil Klasifikasi Data Uji 25% (Metode *Oversampling*)..... 64
Lampiran 15. Hasil Klasifikasi Data Uji 20% (Metode *Oversampling*)..... 64
Lampiran 16. Hasil Klasifikasi Data Uji 10% (Metode *Oversampling*)..... 64
Lampiran 17. *Screenshot* Tampilan Aplikasi..... 64