TUGAS AKHIR – KS141501

PREDIKSI DIABETES BERDASARKAN FAKTOR RISIKO *BEHAVIORAL* MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACIHINE*

PREDICTION OF DIABETES BASED ON BEHAVIORAL RISK FACTOR USING SUPPORT VECTOR MACHINE AGORITHM

ANNISA NURLAILY
NRP 0521 11 4000 0123

Dosen Pembimbing
Renny Pradina Kusumawardani S.T., M.T. SCJP

Departemen Sistem Informasi
Fakultas Teknologi Informasi dan Komunikasi
Institut Teknologi Sepuluh Nopember
Surabaya 2018

TUGAS AKHIR – KS141501

PREDIKSI DIABETES MENGGUNAKAN FAKTOR RISIKO *BEHAVIORAL* MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE*

ANNISA NURLAILY
NRP 05211140000123

Dosen Pembimbing
Renny Pradina Kusumawardani S.T., M.T. SCJP

Departemen Sistem Informasi
Fakultas Teknologi Informasi dan Komunikasi
Institut Teknologi Sepuluh Nopember
Surabaya 2018

TUGAS AKHIR – KS141501

PREDICTION OF DIABETES BASED ON BEHAVIORAL RISK FACTOR USING SUPPORT VECTOR MACHINE ALGORTIHM

ANNISA NURLAILY
NRP 05211140000123

Supervisor
Renny Pradina Kusumawardani S.T., M.T. SCJP

Departement of Information Systems
Faculty of Information Technology and Communication
Institut Teknologi Sepuluh Nopember
Surabaya 2018

LEMBAR PENGESAHAN

PREDIKSI DIABETES BERDASARKAN FAKTOR RISIKO *BEHAVIORAL* MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE*

TUGAS AKHIR
Disusun Untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer
pada
Departemen Sistem Informasi
Fakultas Teknologi Informasi dan Komunikasi
Institut Teknologi Sepuluh Nopember

Oleh:

ANNISA NURLAILY
NRP. 0521 11 4000 0123

Surabaya, 25 Juli 2018

KEPALA
DEPARTEMEN SISTEM INFORMASI

Dr. Ir. Aris Tjahyanto, M.Kom.
NIP 19650310 199102 1 001

LEMBAR PERSETUJUAN

PREDIKSI DIABETES BERDASARKAN FAKTOR RISIKO *BEHAVIORAL* MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE*

Disusun Untuk Memenuhi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer
pada
Departemen Sistem Informasi
Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember

Oleh:

ANNISA NURLAILY
NRP. 0521 11 4000 0123

Disetujui Tim Penguji: Tanggal Ujian: 20 Juli 2018
Periode Wisuda: September 2018

Renny Pradina K., S.T., M.T., SCJP **(Pembimbing I)**

Faizal Johan Atletiko, S.Kom., M.T. **(Penguji I)**

Nur Aini Rakhmawati, S.Kom, M.Sc.Eng, Ph.D **(Penguji II)**

PREDIKSI DIABETES BERDASARKAN FAKTOR RISIKO *BEHAVIORAL* MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE*

Nama Mahasiswa : Annisa Nurlaily

NRP : 0521 11 4000 0123

Jurusan : Sistem Informasi FTIK-ITS

Pembimbing 1 : Renny Pradina Kusumawardani S.T., M.T., SCJP

ABSTRAK

Diabetes merupakan salah satu penyakit yang turut andil dalam penyebab peningkatan jumlah kematian di dunia. Hal ini bisa dilihat dari pesatnya perkembangan angka penyandang diabetes yang naik hampir 2 kali lipat sejak tahun 1980 dari 4,7% menjadi 8,5% di tahun 2014. Di Indonesia, kasus diabetes juga sering dan umum dijumpai. Data yang didapatkan dari hasil penetilian yang dilakukan pusat-pusat diabetes, sekitar tahun 1980 prevalensi diabetes pada penduduk usia 15 tahun ke-atas sebesar 1,5-2,3% dengan kecenderungan di daerah pedesaan lebih rendah dibandingkan daerah perkotaan.

Dengan fakta dan urgensi seperti ini, banyak sudah tindakan-tindakan preventif yang dilakukan untuk dapat menekan pertumbuhan diabetes secara global. Tidak hanya berfokus pada analisi faktor risiko yang berhubunan dengan data medis, analisi faktor risko yang berbungan dengan gaya dan pola hidup juga sudah banyak ditemukan. Diabetes, selain memiliki beberapa faktor risiko yang bisa dikelompokkan menjadi faktor risiko yang tidak dapat dimodifikasi seperti ras dan etnik; umur; jenis kelamin; serta riwayat penyakit keluarga, juga mempunyai faktor risiko yang dapat dimodifikasi seperti berat badan yang berlebih; obesitas; kurangnya aktivitas fisik; serta frekuensi merokok.

Terdapat beberapa penelitian sebelumnya yang telah dilakukan untuk melakukan analisa terhadap faktor risiko diabetes menggunakan algoritma data mining Linear Regression. Hasilnya ditemukan variable yang berpengaruh secara signifikan pada kasus diabetes yang tergolong kedalam faktor risiko behavioral. Oleh karena itu, mengacu pada penelitian sebelumnya, pada penelitian ini akan dilakukan analisis prediksi diabetes berdasarkan beberapa faktor risiko behavioral terkait seperti obesitas, kelebihan berat badan (overweight), umur, konsumsi alcohol, tekanan darah, jenis kelamin, dan ras menggunakan data yang disediakan oleh CDC BFRSS yakni sebuah lembaga survey di Amerika yang secara rutin melakukan wawancara kepada warga Amerika mengenai gaya hidup dan status penyakit kronis khususnya pada tahun 2016 dengan mengunakan metode algoritma Support Vector Machine (SVM). SVM adalah salah satu machine learning dengan model supervised learning dimana algortima ini digunakan penganalisaan data untuk klasifikasi dan dan analisis regresi.

Hasil dari penelitian ini berupa nilai hasil pemodelan prediksi dengan metric pengukuran yang paling baik. Serta perbandingan penggunaan algoritma dengan hasil penelitian sebelumnya, dimana ternyata algoritma SVM untuk menganalisis model data BFRSS 2016 tidak terlalu optimal.

Kata kunci: prediksi, support vector machine, diabetes, faktor risiko behavioral

PREDICTION OF DIABETES BASED ON BEHAVIORAL RISK FACTOR USING SUPPORT VECTOR MACHINE ALGORITHM

Student Name	**: Annisa Nurlaily**
NRP	**: 0521 11 400 0123**
Department	**: Sistem Informasi FTIf-ITS**
Supervisor 1	**: Renny Pradina Kusumawardani S.T., M.T., SCJP**

ABSTRACT

Diabetes is one of the diseases that contribute to the cause of the increasing number of deaths in the world. This can be seen from the rapid development of the number of people with diabetes who rose almost 2-times fold since 1980 from 4.7% to 8.5% in 2014. In Indonesia, diabetes cases are also common. Data obtained from the results of reseacrh conducted by diabetes centers, around 1980 the prevalence of diabetes in the population aged 15 years to the maximum of 1.5-2.3% with the trend in rural areas lower than urban areas.

With these facts and urgency, many preventive measures have been taken to suppress the growth of diabetes globally. Not only focusing on risk factor analysis that is related to medical data, risk factor analysis that is related to style and lifestyle has also been found. Diabetes, in addition to have several risk factors that can be grouped into unmodified risk factors such as race and ethnicity; age; gender; as well as family history of the disease, also have modifiable risk factors such as excess weight; obesity; lack of physical activity; as well as the frequency of smoking.

There have been several previous studies that have been done to analyze the risk factors for diabetes using the Linear Regression data mining algorithm. The results found that variables significantly influence in cases of diabetes belonging to behavioral risk factors. Therefore, referring to previous research, this study will analyze diabetes prediction based on several behavioral related risk factors such as obesity, overweight, age, alcohol consumption, blood pressure, sex, and race using the data provided by CDC BFRSS, a survey agency in America that routinely interviews Americans about lifestyle and chronic disease status especially in 2016 using the Support Vector Machine (SVM) algorithm method. SVM is one of machine learning with supervised learning model where this algorithm is used for analyzing data for classification and and regression analysis.

The result of this research is the result of prediction model with the best metric measurement. And comparison of the use of the algorithm with result of previous research, where turns SVM algorithm to analyze model data BFRSS 2016 is not too optimal.

Keywords**: **predicition, support vetor machine, diabetes, risk behavioral factor

KATA PENGANTAR

Puji dan syukur penulis tuturkan ke hadirat Allah SWT, Tuhan Semesta Alam yang telah memberikan kekuatan dan hidayah-Nya kepada penulis sehingga penulis mendapatkan kelancaran dalam menyelesaikan tugas akhir dengan judul:

PREDIKSI DIABETES BERDASARKAN FAKTOR RISIKO *BEHAVIORAL* MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE*

yang merupakan salah satu syarat kelulusan pada Jurusan Sistem Informasi, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember Surabaya.

Terima kasih penulis sampaikan kepada pihak-pihak yang telah mendukung, memberikan saran, motivasi, semangat, dan bantuan baik berupa materiil maupun moril demi tercapainya tujuan pembuatan tugas akhir ini. Tugas akhir ini tidak akan pernah terwujud tanpa bantuan dan dukungan dari berbagai pihak yang sudah melauangkan waktu, tenaga dan pikirannya. Secara khusus penulis akan menyampaikan ucapan terima kasih yang sebanyak-banyaknya kepada:

1) Orang tua dan keluarga penulis yang telah memberikan motivasi, semangat, keyakinan, kasih sayang serta doa sehingga penulis mampu menyelesaikan pendidikan S1 ini dengan baik.
2) Ibu Renny Pradina Kusumawardani ST., MT. selaku dosen pembimbing yang telah dengan sabar dan telaten memberikan ilmu, petunjuk, dan motivasi sehingga penulis dapat menyelesaikan Tugas Akhir ini.
3) Bapak Faizal Johan Atletiko, S.Kom, M.T, dan Ibu Nur Aini Rakhmawati, S.Kom., M.Sc., Eng. Ph.D selaku dosen penguji yang telah memberikan masukan-masukan guna menyempurnakan Tugas Akhir ini.

4) Seluruh Dosen Departemen Sistem Informasi ITS yang telah memberikan ilmu pengetahuan yang bermanfaat dan pengalaman yang berharga bagi penulis
5) Serta seluruh pihak-pihak lain yang tidak dapat disebutkan satu per satu yang telah banyak membantu penulis selama perkuliahan hingga dapat menyelesaikan tugas akhir ini.

Penulis sadar bahwa Tugas Akhir ini masih jauh dari kata sempurna, sehingga saran dan kritik yang membangun dari pembaca merupakan *feedback* yang berarti untuk perbaikan ke depan. Semoga Tugas Akhir ini dapat bermanfaat bagi perkembangan ilmu pengetahuan dan semua pihak.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR KODE

BAB I
PENDAHULUAN

Pada bab ini akan dibahas mengenai latar belakang pengerjaan tugas akhir, rumusan permasalahan, batasan permasalahan, tujuan pengerjaan dan juga manfaat pengerjaan tugas akhir an.

1.1 Latar Belakang

Diabetes merupakan salah satu penyakit yang turut andil dalam penyebab peningkatan jumlah kematian di dunia. Hal ini bisa dilihat dari pesatnya perkembangan angka penyandang diabetes yang naik hampir 2 kali lipat sejak tahun 1980 dari 4,7% menjadi 8,5% di tahun 2014 [**1**]. Selain sebagai penyebab langsung kematian, diabetes juga berperan sebagai pemicu perkembangan penyakit lain seperti kebutaan, gagal ginjal, serangan jantung, dan stroke. Diabetes dikategorikan menjadi 2 yakni diabetes tipe 1 dan diabetes tipe 2. Diabetes tipe 1 ditandai dengan berkurangnya produktivitas insulin dalam tubuh akibat sistem kekebalan tubuh yang menyerang dan merusak cell yang memproduksi insulin, sedangkan diabetes tipe 2 disebabkan oleh penggunaan insulin yang tidak efektif oleh tubuh, pada umumnya disebabkan oleh kurangnya aktivitas fisik dan kelebihan berat badan. Insulin sendiri adalah hormon yang mengatur keseimbangan gula dalam darah.

Di Indonesia, kasus diabetes juga sering dan umum dijumpai. Data yang didapatkan dari hasil penetilian yang dilakukan pusat-pusat diabetes, sekitar tahun 1980 prevalensi diabetes pada penduduk usia 15 tahun ke-atas sebesar 1,5-2,3% dengan kecenderungan di daerah pedesaan lebih rendah dibandingkan daerah perkotaan. Persentase angka ini terus meningkat berdasarkan data yang diperoleh Survei Kesehatan Rumah Tangga (SKRT) pada 2001 yakni sebesar 7,5% pada penduduk usia 25-64 tahun di Jawa-Bali. Selain itu pada 2007 dan 2013, Riset Kesehatan Dasar (Riskedas) melakukan survei wawancara untuk menghitung proporsi penderita diabetes

pada usia 15 tahun ke-atas, dimana ditemukan fakta bahwa proporsi diabetes tahun 2013 meningkat hampir dua kali lipat dibandingkan tahun 2007. Sayangnya jumlah proporsi yang besar ini belum meliputi penderita diabetes yang belum terdiagnosis, maka jika akan dilakukan survei secara menyeluruh dan merata kemungkinan akan didapatkan jumlah proporsi diabetes lebih besar lagi [**2**].

Diabetes memiliki beberapa faktor risiko yang bisa dikelompokkan menjadi faktor risiko yang tidak dapat dimodifikasi seperti ras dan etnik; umur; jenis kelamin; serta riwayat penyakit keluarga, dan faktor risiko yang dapat dimodifikasi seperti berat badan yang berlebih; obesitas; kurangnya aktivitas fisik, hipertensi; riwayat Toleransi Glukosa Terganggu (TGT) atau Gula Darah Puasa terganggu (GDP terganggu); dan merokok. Di Amerika Serikat, terdapat sebuah lembaga bernama The Center for Disease Control and Prevention (CDC) yang menyediakan data-daya yang bisa diakses oleh publik mengenai berbagai faktor risiko penyakit yang dikenal dengan The Behavioral Risk Factor Surveillance System (BRFSS). BRFSS menghimpun data faktor risiko tahunan dari berbagai wilayah Negara bagian di Amerika Serikat dengan melakukan survey atau wawancara melalui telepon sejak 1984 hingga sekarang. Informasi yang dihimpun berupa kondisi kesehatan individu dan faktor risiko yang berhubungan dengan perilaku seperti umur, jenis kelamin, ras atau etnik, merokok, tekanan darah, dan hal lain yang berkaitan. Saat ini karena sistem survey atau pengamatan yang dilakukan oleh BRFSS sangat memadai dan membantu banyak pihak untuk mengembangkan analisa guna meningkatkan promosi kesehatan yang lebih baik, banyak negara yang ingin mengadopsi sistem ini diantaranya Australia, Brasil, Canada, Korea, dan lain sebagainya [**3**].

Terdapat beberapa penelitian sebelumnya yang telah dilakukan untuk melakukan analisa terhadap faktor risiko diabetes menggunakan bermacam algoritma *data mining* salah satunya adalah *Support Vector Machine* (SVM). SVM adalah salah

satu *machine learning* dengan model *supervised learning* dimana algortima ini digunakan penganalisaan data untuk klasifikasi dan dan analisi regresi. Oleh karena itu pada penelitian ini akan dilakukan analisis prediksi diabetes berdasarkan beberapa faktor risiko behavioral terkait seperti obesitas, kelebihan berat badan (overweight), umur, konsumsi alcohol, tekanan darah, jenis kelamin, dan ras menggunakan data yang disediakan oleh CDC pada tahun 2016 dengan mengunakan metode algoritma Support Vector Machine (SVC).

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan di atas, rumusan masalah yang akan diselesaikan pada penelitian ini adalah bagaimana melakukan prediksi diabetes berdasarkan data-data faktor risiko behavioral terkait dengan menggunakan algoritma Support Vector Machine beserta hasil prediksi dan akurasinya

1.3 Batasan Maslaah

Pada penelitian kali ini, berikut adalah beberapa batasan masalah yang diterapkan:

1. Dataset yang digunakan untuk training dan testing adalah data yang disediakan oleh BRFSS pada tahun 2016 [**4**].
2. Terdapat 10 atribut yang digunakan dalam prediksi kali ini yakni *State, Sex, Age, Ethnicity, Overweight, Obesity, Smoke, Alcohol, Physical Activity,* dan *Diabetes.*

1.4 Tujuan

Tujuan dari pengerjaan Tugas Akhir ini adalah:

1. Menerapkan prediksi dan mengetahui hasil prediksi diabetes menggunakan algoritma SVM.

2. Mengetahui berapa nilai C dan Gamma yang terbaik untuk menghasilkan akurasi prediksi yang tinggi.

1.5 Manfaat

Tugas Akhir ini diharapkan dapat memberikan manfaat yaitu:

Bagi penulis

Penulis dapat mengimplementasikan teknik *data mining* menggunakan algoritma SVM pada dataset tertentu.

Bagi masayarakat

Mengetahui faktor risiko behavioral yang signifikan terhadap kasus diabetes pada umumnya.

1.6 Relevansi

Relevansi dari pengerjaan tugas akhir ini adalah untuk syarat kelulusan dimana melakukan implementasi teori keilmuan yang di dapatkan pada mata kuliah Penggalian Data dan Analitika Bisnis dan Sistem Keputusan.

BAB II
TINJAUAN PUSTAKA

Untuk dapat mengetahui wawsan dan gambaran secara umum mengenai beberapa hal dan topik yang akan dibahas dalam tugas akhir ini, berikut ini adalah penjelasan mengenai penelitian sebelumnya yang telah dilakukan yang dijadikan bahan acuan dan rujukan dalam pengerjaan tugas akhir serta beberapa dasar teori yang dapat membantu mempermudah proses pemahaman proses hasil pembahasan tugas akhir ini.

2.1 Penelitian Sebelumnya

Beberapa penelitian sebelumnya yang terkait dengan tugas akhir ini adalah sebagai berikut.

Tabel 2.1 Penelitian Sebelumnya

No	Judul Penelitian - Tahun	Identitas Peneliti	Kesimpulan	Sumber
1.	*Examining Disease Risk Factor by Mining Publicly Available Information - 2013*	Jay Pederse n, Fangyao Liy, Harry Ngondo	Berdasarkan hasil penelitian, dihasilkan kesimpulan bahwa model yang digunakan mengindikasikan bahwa faktor risiko yang paling penting berdasarkan koefisiensi variable yng digunakan adalah variable obesitas.	[**5**]

No	Judul Penelitian - Tahun	Identitas Peneliti	Kesimpulan	Sumber
2.	*Prevalence of obesity, diabetes, and obesity-related health factors - 2001*	Mokdad AH, Ford ES, Bowman BA, et al	Berdasarkan hasil penelitian, kesimpulan yang didapatkan adalah bahwa terjadi peningkatan obesitas dan diabetes warga Amerika Serikat pada kedua sample jenis kelamin, umur, semua suku dan ras, semua tingkat pendidikan, dan semua tingkat perokok. Serta data bahwa obesitas berpengaruh atau berasosiasi dengan beberapa kebanyakan faktor risiko kesehatan.	[**6**]

2.2 Dasar Teori

2.2.1 Diabetes

Diabetes merupakan salah satu faktor besar meningkatnya jumlah angka kematian dunia, bahkan WHO memprediksi

bahwa diabetes akan menjadi faktor pembunuh langsung nomor 6 di dunia pada tahun 2030 mendatang. Diabetes dikategorikan menjadi 2 tipe yakni diabetes tipe 1 dan diabetes tipe 2.

Diabetes tipe 2 secara signifikan terus meningkat jumlahnya namun disisi lain sebenarnya sangat mungkin untuk dilakukan pencegahan. Berdasarkan Centers for Disease Control and Prevention (CDC), diabetes tipe 2 diderita oleh 90 hingga 95 persen pasien yang didiagnosa terkena diabetes pada orang dewasa. Secara umum berikut adalah gambaran dari kasus diabetes:

1. Penelitian membuktikan bahwa 1 dari 3 orang dewasa memiliki pre-diabetes, dan pada kasus ini 9 dari 10 orang tidak mengetahui bahwa mereka mempunyai pre-diabetes.
2. 29,1 juta orang di Amerika Serikat memiliki diabetes namun 8,1 juta diantaranya kemungkinan tidak terdiagnosa dan tidak waspada terhadap konsidi mereka.
3. Sekitar 1,4 juta kasus baru diabetes didiagnosa di Amerika Serikat setiap tahunnya.
4. Lebih dari setiap 10 orang dewasa dengan umur 20 tahun ke atas yang memiliki diabetes. Untuk lansia yang berumur 65 tahun ke atas juga memiliki kecenderungan peningkatan risiko diabetes dengan ratio kemungkinan 1 dari 4 orang.

2.2.2 Faktor Risiko

Faktor risiko diabetes terutama diabetes tipe 2 termasuk didalamnya adalah mengenai pola dan gaya hidup, namun hal ini bisa dikurangi atau bahkan dihilangkan seiring berjalannya waktu dan usaha yang dikeluarkan. Dalam kasus diabetes, kebanyakan jenis kelamin pria lebih beresiko terkena ketimbang wanita. Beberapa faktor dari kemungkinan ini berhubungan dengan faktor gaya hidup, berat badan dan

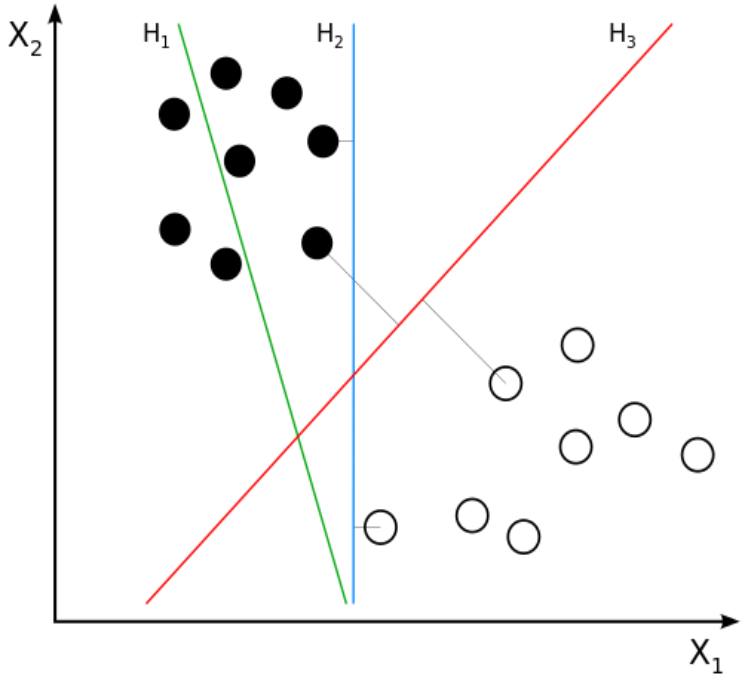
obesitas. Berikut ini adalah beberapa faktor risiko diabetes yang signifikan:

1. Umur
   Risiko diabetes seakin meningkat dengan bertambahnya usia. Walaupun terdapat kasus diabetes tipe 2 pada anak-anak namun hal ini pada umumnya dikarenakan adanya factor kelebihan berat badan pada usia muda.
2. Kelebihan Berat Badan Obesitas
   Berat badan yang berlebihan bisa memicu timbulnya obesitas yang merupakan salah satu faktor timbulnya berbagai macam penyakit kronis termasuk diabetes.
3. Riwayat Keluarga
   Orang dengan riwayat keluarga pernah didiagnosa diabetes meningkatkan kemungkinan terkena diabetes daripada orang yang tidak mempunyai riwayat kelaurga pasien diabetes.
4. Ras atau Suku
   Ras dan suku tertentu dapat mempunyai kecenderungan tingkat rata-rata yang tinggi untuk kasus pre-diabetes dan diabetes tipe 2. Sebagai contoh penduduk asli Amerika lebih umum terkena diabetes tipe ketimbang ras Kaukasian.
5. Kurangnya Aktivitas Fisik
   Kurangnya aktivitas fisik mempengaruhi kinerja tubuh secara keseluruhan sehingga berdampak pada kinerja tubuh yang tidak optimal.

2.2.3 *Support Vector Machine*

Support Vector Machine (SVM) adalah salah satu metode supervised learning yang digunakan untuk kebutuhan klasifikasi, regresi dan pendetesian outlier. SVM juga dikenal sebagai teknik pembelajaran mesin (machine learning) paling

mutakhir setelah pembelajaran mesin sebelumnya yang dikenal sebagai Neural Network (NN). Baik SVM maupun NN tersebut telah berhasil digunakan dalam pengenalan pola. Pembelajaran dilakukan dengan menggunakan pasangan data input dan data output berupa sasaran yang diinginkan. Pembelajaran dengan cara ini disebut dengan pembelajaran terarah (*supervised learning*). Dengan pembelajaran terarah ini akan diperoleh fungsi yang menggambarkan bentuk ketergantungan input dan outputnya. Selanjutnya, diharapkan fungsi yang diperoleh mempunyai kemampuan generalisasi yang baik, dalam arti bahwa fungsi tersebut dapat digunakan untuk data input di luar data pembelajaran. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari hyperplane terbaik yang berfungsi sebagai pemisah dua buah class pada input space Gambar 2.1.



Gambar 2.1. Hyperplane pada SVM

Karakteristik SVM :

1. Secara prinsip SVM adalah linear classifier.
2. *Pattern recognition* dilakukan dengan mentransformasikan data pada input space ke ruang yang berdimensi lebih tinggi, dan optimisasi dilakukan pada ruang vector yang baru tersebut. Hal ini membedakan SVM dari *solusi pattern recognition* pada umumnya, yang melakukan optimisasi parameter pada ruang hasil transformasi yang berdimensi lebih rendah daripada dimensi *input space*.
3. Menerapkan strategi *Structural Risk Minimization* (SRM).
4. Prinsip kerja SVM pada dasarnya hanya mampu menangani klasifikasi dua class.

Beberapa keunggulan dalam penggunaan SVM adalah:

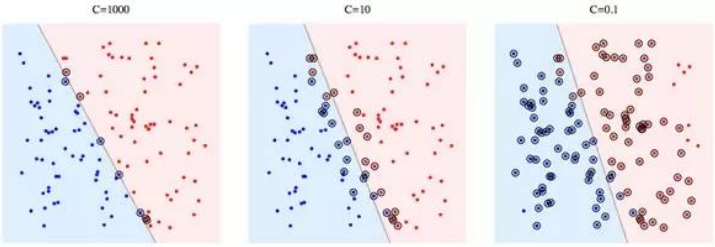
1. Efektif digunakan dalam ruang dimensi yang tinggi.
2. Cukup efektif jika digunakan dalam kasus dimana jumlah dari dimesi lebih besar dari jumlah sample yang digunakan.
3. Menggunakan subset dari training points yang berada pada fungsi decision yang dinamakan dengan *support vectors*.
4. Serbaguna: fungsi Kernel yang berbeda dapat dispesifikasikan sesuai dengan fungsi decision yang dipilih.

Beberapa kekurangan dalam penggunaan SVM adalah:

1. Jika jumlah dari fitur jauh lebih besar daripada jumlah dari sample yang digunakan maka langkah over-fitting dalam pemilihan fungsi Kernel dan regularisasi perlu untuk dilakukan.
2. SVM tidak secara langung menghasilkan estimasi kemungkinan, maka dari itu untuk perhitungannya perlu digunakan meotde five-fold cross validation.

2.2.3.1 Kernel

Algoritma SVM menggunakan serangkaian fungsi matematika yang didefinisikan sebagai kernel. Fungsi dari kerel adalah untuk mengambil data sebagai input dan mentrasformasikannya menjadi bentu yang dibutuhkan. Penggunaan algoritma yang berbeda dalam SVM juga akan mempengaruhi tipe dari fungsi kernel. Sebagai contoh, *linear; nonlinear, polynominal, radial basis function (RBF); dan sigmoid.* Fungsi kernel yang paling banyak digunakan adalah fungsi RBF. Pada fungsi kernel nonlinear dan RBF, digunakan parameter C dan Gamma. C adalah parameter yang digunakan untuk fungsi *soft margin cost,* dimana mengatur pengaruh dari masing-masing *support vector;* proses ini melibatkan penukaran *error penalty* untuk kestabilan. Sedangkan Gamma adalah parameter bebas dari fungsi Gaussian radial basis. Gamma dengan nilai kecil berarti Gaussian dengan varian yang besar sehingga pengaruh dari nilai x_i semkan besar. Namun jika nilai Gamma semakin besar, makan semakin kecil nilai varian maka *support vector* tidak akan mempunyai pengaruh yang luas lagi.



Gambar 2.2. Contoh SVM dengan beberapa nilai C

2.2.3.2 Performance Measurement Metrics

1. *Accuracy*
   Secara umum, metric akurasi menghitung rasio dari jumlah prediksi yang benar dari keseluruhan total angka instansi yang dievaluasi.

$$\frac{tp + tn}{tp + fp + tn + fn}$$

2. *Precision*
   Presisi digunakan untuk pola positif yang diprediksi secara benar dari total seluruh pola prediksi dikelas positif.

$$\frac{tp}{tp + fp}$$

3. *Recall*
   Recall digunakan untuk menghitung fraksi pola positif yang terklasifikasi secara benar.

$$\frac{tp}{tp + tn}$$

4. *F-Measure*
   F-Measure, metrik ini merepresentasikan nilai tengah (mean) yang seimbang antara nilai *recall* dan nilai presisi.

$$\frac{2 * p * r}{p + r}$$

2.2.4 *Python*

Python adalah bahasa pemrograman yang bersifat open source. *Python* telah digunakan untuk mengembangkan berbagai macam perangkat lunak, seperti *internet scripting, systems programming, user interface, product customization, numberic programming,* dll.

Pemrograman bahasa *python* adalah pemrogram gratis atau freeware, sehingga dapat dikembangkan, dan tidak ada batasan dalam penyalinannya dan distribusi. Terdapat beberapa pelayanan yang disediakan lengkap dengan *source code, debugger* dan *profiler, interface*, fungsi sistem, GUI, dan basis datanya. *Python* tersedia untuk berbagai sistem operasi seperti Unix (linux), PCs (Dos, Windows, OS/2), Machintosh, dsb.

Bahasa pemrograman *Python* memiliki beberapa fitur yang dapat digunakan oleh pengembang perangkat lunak. Berikut adalah beberapa fitur yang ada pada bahasa pemrograman *Python* :

- *Multi Paradigm Design*
- *Open Source*
- *Simplicity*
- *Library Support*
- *Portability*
- *Extendable*
- *Scalability*

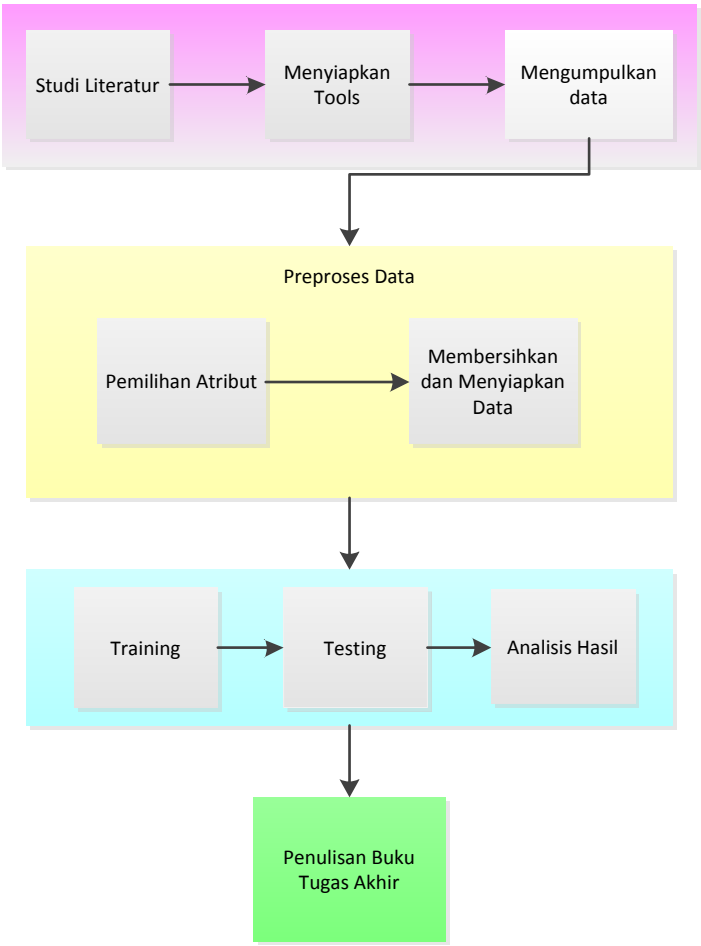
Kelebihan dari *Python* :

- Tidak ada deklarasi tipe data yang merumitkan sehingga program menjadi lebih sederhana, singkat, dan fleksibel
- Manajemen memori otomatis
- Pemrograman berorientasi objek
- Pelekatan dan perluasan dalam C
- Terdapat kelas modul, eksepsi, sehingga terdapat dukungan pemrograman skala besar secara modular
- Interaktif, dinamis, dan alamiah
- Portabilitas secara luas seperti pemrograman antarplatform tanpa ports

Halaman ini sengaja dikosongkan.

BAB III
METODOLOGI

Pada bab ini akan dijelaskan mengenai gambaran metode dan alur pengerjaan tugas akhir. Berikut ini merupakan alur pengerjaan tugas akhir Gambar 3.1.



Gambar 3.1 Bagan Metodologi

3.1 Business Understanding

Dalam proses studi literature, hal yang dilakukan adalah dengan meninjau dan memahami dasar-dasar teori yang berkaitan dengan permasalahan yang akan diselesaikan pada tugas akhir ini. Beberapa teori tersebut adalah mengenai teknik data mining untuk melakukan prediksi dengan menggunakan algoritma *Support Vector Machine*. Selain itu juga memahami mengenai topic bahasan yakni diabetes dan faktor risiko terkait. Penelitian-penelitian sebelumnya yang pernah dilakukan mengenai topik terkait juga menjadi bahan studi literatur untuk membantu proses pemahaman topik lebih jauh yang nantinya bisa dijadikan bahan rujukan jika terjadi kebingungan dalam proses pengerjaan.

3.1.1 Menyiapkan Tools

Setelah melakukan studi literature dengan seksama, maka langkah selanjutnya yang dilakukan akan menyiapkan tools yang akan digunakan untuk melakukan pengolahan data dan proses prediksi seperti mempersiapkan Jupyter Notebook sebagai tools yang mendukung bahasa pemograman Python.

3.2 Data Understanding

3.2.1 Pengumpulan Data

Pada proses ini, data diperoleh dari publikasi *Center for Disease Control and Prevention* mengenai data *Behavioral Factor Risk Surveillance System* (BFRSS) pada tahun 2016. Telah disediakan beberapa tipe data file yang dapat diunduh diantaranya tipe data file ASCII dan tipe data file dengan *SAS Transport Format*. Selain menyiapkan file data untuk bahan dataset, terdapat juga beberapa informasi mengenai data survey beserta dokumentasinya yang perlu diperhatikan pada proses ini untuk bisa memahami fitur data dan tipe data yang disediakan seperti *2016 BRFSS Overview* dan *BFRSS 2016 Codebook*. Dalam BRFSS Overview terdapat informasi mengenai desain kuisioner seperti standar pertanyaan bagaiamana yang digunakan serta topik apa saja yang menjadi bahan pertanyaan. Sedangkan dalam BRFSS Codebook

terdapat informasi mengenai sejumlah pertanyaan wawancara yang berhasil dihimpun dan dikategorikan kedalam label tertentu serta penamaan variable pada dataset yang nantinya akan sangat berguna sebagai rujukan dalam proses pemilihan atribut dataset.

3.3 Data Preparation

3.3.1 Pemilihan Atribut

Pada tahap ini, berdasarkan hasil studi literatur mengenai faktor-faktor risiko diabetes pada penelitan-penelitian yang telah lebih dulu dilakukan, terdapat beberapa faktor signifikan yang mempengaruhi risiko diabetes yakni seperti usia, jenis kelamin, informasi mengenai berat badan dan obesitas, dan gaya serta pola hidup seperti merokok; konsumsi alkohol; dan

3.3.2 Membersihkan dan Menyiapkan Data

Pada tahap ini yang dilakukan dalam membersihkan dan menyiapkan data adalah melakukan *treatment* pada *missing value* dengan menghapus baris data yang memiliki *missing value.* Selain itu jika terdapat beberapa tipe dari data yang tidak sesuai dengan apa yang dibutuhkan, maka akan dilakukan penyesuaian. Dalam hal ini tipe data yang digunakan untuk proses prediksi adalah tipe ordinal.

3.4 Modelling

3.4.1 Training Data

Pada tahap ini proses untuk menghasilkan data training yang pertama adalah memastikan bahwa dataset yang sudah disiapkan memiliki *class label* yang diinginkan. Setelah itu proses *training data* akan dilakukan dengan metode *cross validation kfold.* Pada tahap ini akan dihasilkan model prediksi untuk melakukan proses *testing*.

3.4.2 Testing

Pada tahap ini setelah didapatkan model prediksi dari *data training* yang telah dilakukan yang kemudian model tersebut akan diterapkan pada dataset test yang sudah ditentukan. Proses ini akan menghasilkan nilai performa prediksi yang telah dilakukan berupa nilai presisi, recall, f-score, dan

3.5 Analisis Hasil Prediksi

Tahap ini adalah tahap melakukan analisis hasil yang didapatkan dari proses prediksi yang telah dilakukan. Nilai akurasi, presisi, *recall*, dan *f-score* akan dibandingkan berdasarkan nilai C dan Gamma yang akan diubah

3.6 Penulisan Buku Tugas Akhir

Tahapan terakhir dari pengerjaan tugas akhir ini adalah pembuatan dokumentsi buku tugas akhir. Buku ini aakan berisi paparan dan penjelasan mengenai latar belakang pengerjaan, tujuan pengerjaan, dokumentasi pengerjaan hingga hasil analisis hasil yang didapatkan.

BAB IV
PERANCANGAN

Pada bab ini akan dipaparkan mengenai perancangan penelitian tugas akhhir.

4.1 Pengambilan Data

Data BFRSS 2016 yang disediakan oleh CDC memiliki tipe data file ASCII dan *SAS Transport Format*, dimana kedua tipe data ini memerlukan program khusus untuk dapat dibuka.

Data tipe ASCII merupakan file berupa teks dimana setiap *byte* merepresentasikan satu karakter berdasarkan kode ASCII. Kebalikan dari file biner, dimana tidak terdapat *one-to-one mapping* antara *bytes* dan karakter. File yang telah diformat dengan program pengolah kata harus disimpan dan ditransmisikan sebagai file biner untuk mempertahankan pemformatan. File ASCII kadang juga disebut sebagai *plain text file* [7]. Pada umumnya, file ASCII berekstensi .ASC atau .TXT dan dapat dibuka menggunakan *text editor* seperti Notepad atau Sublime Text, namun untuk file dengan jumlah data yang besar seperti file BFRSS 2016, perlu dipertimbangkan pemilihan *text editor* yang tepat untuk dapat membuka dan melakukan perubahan pada data.

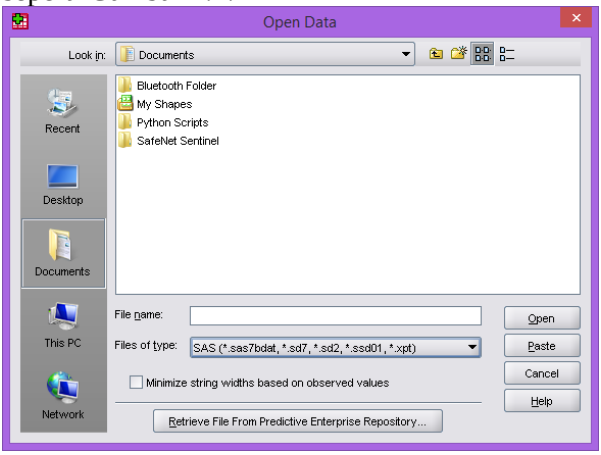
Tipe data yang kedua yakni *SAS Transport Format* dimana dataset diekspor dari SAS V9.3 dengan ekstensi file .XPT dengan atribut sebanyak 275 atribut. Format data ini dapat dibuka menggunakan SAS Universal Viewer, namun terdapat juga beberapa program yang dapat digunakan untuk mengimpor data seperti STATA atau SPSS.

Dalam penelitian kali ini, file yang diguankan adalah file BFRSS 2016 dengan ekstensi .XPT dan proses impor data menggunakan SPSS untuk bisa dilakukan penyimpanan dataset dengan format .CSV.

4.1.1 Impor Data Menggunakan SPSS

Proses impor data menggunakan SPSS adalah sebagai berikut:

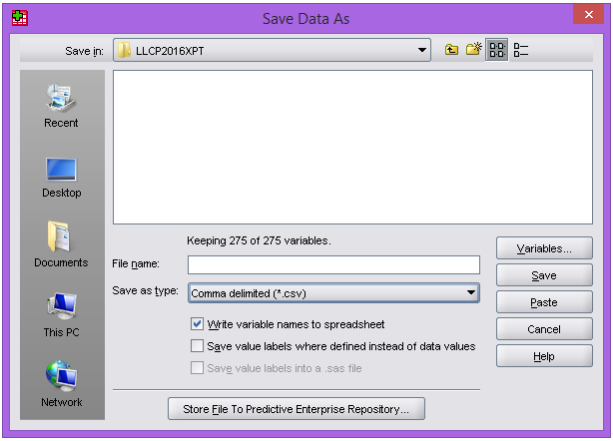
1. Buka SPSS
   Pada window utama SPSS pilih tab **File** > **Open** > **Data**, lalu akan muncul window seperti Gambar 4.1.



Gambar 4.1 Window Open Data

2. Pilih tipe data
   Pilih tipe data sebagai SAS dan pilih nama file berekstensi .XPT yang akan dibuka lalu pilih **Open**. Selanjutnya SPSS akan menjalankan perintah yang diberikan. Proses impor ini akan membutuhkan beberapa waktu dikarenakan jumlah data yang besar.

3. Expor data ke format .CSV
   Setelah data telah terbuka, maka langkah selanjutnya adalah export data ke dalam format .CSV (Gambar 4.2). Pada tab **File** > **Save As** > **pilih format data menjadi Comma Delimited (.csv)** > **isi nama file** > **Save**.



Gambar 4.2 Window Save Data As

4.2 Pemilihan Atribut

Berdasarkan hasil penelitian sebelumnya mengenai prevalensi obesitas dan kelebihan berat badan terhadap diabetes menunjukkan bahwa kedua hal tersebut berpengaruh positif dan signifikan terhadap diabetes. Penelitian sebelum lainnya mengenai faktor risiko diabetes dimana menggunakan dataset BFRSS tahun 2002, menunjukkan bahwa terdapat sejumlah atribut yang berpengaruh cukup signifikan terhadap kasus diabetes selain obesitas dan kelebihan berat badan seperti jumlah konsumsi alkohol, status perokok, dan intensitas aktivitas fisik yang dilakukan. Oleh karena itu, merujuk pada dua hasil penelitian sebelumnya, dalam tahap ini akan dilakukan ekstraksi variable atribut yang serupa yakni:

1. *DISPOSITION* – menunjukkan bahwa survei dilakukan secara menyeluruh (tidak ada pertanyaan yang dilewatkan untuk ditanyakan) atau tidak. Berikut ini adalah rincian dari value dari variable Disposition (**Error! Reference source not found.**).

Tabel 4.1 Variable Disposition

Label: Final Disposition Section Name: Record Identification Section Number: 0 Question Number: 14 Column: 32-35 Type of Variable: Num SAS Variable Name: DISPCODE Question Prologue: Question: Final Disposition				
Value	Value Label	Frequency	Percentage	Weighted Percentage
1100	Completed Interview	421,192	86.61	79.19
1200	Partial Complete Interview	65,111	13.39	20.81

2. *STATE* – menunjukkan informasi di wilayah bagian Negara Amerika Serikat mana responden tinggal. Berikut adalah rincian value dari variable State (**Error! Reference source not found.**).

Tabel 4.2 Variable State

Label: State FIPS Code Section Name: Record Identification Section Number: 0 Question Number: 1 Type of Variable: Num SAS Variable Name: _STATE Question Prologue: Question: Stat FIPS Code				
Value	Value Label	Frequency	Percentage	Weighted Percentage

1	Alabama	7,01	1.45	1.49
2	Alaska	2,914	0.60	0.22
4	Arizona	10,952	2.25	2.09
5	Arkansas	5,298	1.09	0.90
6	California	11,393	2.34	11.98
…	…	…	…	…
55	Wisconsin	5,271	1.08	1.77
56	Wyoming	4,497	0.92	0.18
66	Guam	1,578	0.32	0.04
72	Puerto Rico	5,794	1.19	1.10
78	Virgin Islands	1,266	0.26	0.03

3. *AGE* – menunjukkan informasi usia dari responden dengan rentang usia antara 18 tahun hingga lebih dari 65 tahun yang dibagi kedalam 6 level rentang. Berikut adalah rincian value dari variable Age (**Error! Reference source not found.**).

Tabel 4.3 Varibale Age

Label: Imputed age in six groups Section Name: Calculated Variables Section Number: 8 Question Number: 14 Column: 1980 Type of Variable: Num SAS Variable Name: _AGE_G Question Prologue: Question: Six-level imputed age category				
Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Age 18 to 24 Notes: 18 <= _IMPAGE <= 24	26,632	5.48	12.66

2	Age 25 to 34 Notes: 25 <= _IMPAGE <= 34	48,518	9.98	17.38
3	Age 35 to 44 Notes: 35 <= _IMPAGE <= 44	54,873	11.28	16.34
4	Age 45 to 54 Notes: 45 <= _IMPAGE <= 54	77,200	15.87	16.94
5	Age 55 to 64 Notes: 55 <= _IMPAGE <= 64	107,247	22.05	16.66
6	Age 65 or older. Notes: _IMPAGE <= 65	171,833	35.33	20.02

4. *SEX* – Tabel 4.4 berisi informasi jenis kelamin responden.

Tabel 4.4 Variable Sex

Label: Respondents Sex Section Name: Demographics Section Number: 8 Question Number: 1 Column: 120 Type of Variable: Num SAS Variable Name: SEX Question Prologue: Question: Indicate sex of respondent.				
Value	Value	Frequency	Percentage	Weighted

	Label			Percentage
1	Male	210,606	43.31	48.66
2	Female	275,631	56.68	51.33
3	Refused	66	0.01	0.01

5. *ETHNICITY* – menunjukkan informasi ras atau suku responden (Tabel 4.5).

Tabel 4.5 Variable Ethnicity

Label: Computed Race-Ethnicity grouping Section Name: Calculated Race Variables Section Number: 8 Question Number: 7 Column: 1971 Type of Variable: Num SAS Variable Name: _RACE Question Prologue: Question: Race/ethnicity categories				
Value	Value Label	Frequency	Percentage	Weighted Percentage
1	White only, non-Hispanic Notes: _HISPANC = 2 and _MRACE = 1	368,048	75.68	61.89
2	Black only, non- Hispanic Notes: _HISPANC = 2 and _MRACE = 1	39,555	8.13	11.59
3	American Indian or Alaskan Native only, non-Hispanic	7,238	1.49	0.94

	Notes: _HISPANC = 2 and _MRACE = 3			

4	Asian only, non-Hispanic Notes: _HISPANC = 2 and _MRACE = 4	10,492	2.16	5.14
5	Native Hawaian or other Pacific Islander only, non-Hispanic Notes: _HISPANC = 2 and _MRACE = 5	1,444	0.30	0.19
6	Other race only, non-Hispanic Notes: _HISPANC = 2 and _MRACE = 6	2,177	0.45	0.35
7	Multiracial, non-Hispanic Notes: _HISPANC = 2 and _MRACE = 7	9,442	1.94	1.43
8	Hispanic Notes: _HISPANC=1	39,224	8.07	16.58

9	Don’t know/Not sure/Refused Notes: _HISPANC = 7 or 9 or _MRACE = 77 or 99 and _HISPANC = 2	8,683	1.79	1.89

6. *ALCOHOL* – berisi informasi apakah responden mengkonsumsi minumal beralkohol pada 30 hari terakhir. Beikut ini adalah rincian value variable Alcohol (Tabel 4.6).

Tabel 4.6 Variable Alcohol

Label: Drink any alcoholic beverages in past 30 days
Section Name: Calculated Variables
Section Number: 11
Question Number: 1
Column: 2005
Type of Variable: Num
SAS Variable Name: DRNKANY5
Question Prologue:
Question: Adults who reported having had at least one drink of alcohol in the past 30 days

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	204,239	43.43	40.14
2	No – Go to Section 09.05 USENOW3	262,670	55.33	58.33
7	Don’t know/Not Sure – Go to Section 09.05 USENOW3	2,007	0.43	0.37

9	Refused – Go to Section 09.05 USENOW3	1,325	0.28	1.15
Blank	Not asked or Missing	16,062		

7. *SMOKE* - berisi informasi apakah responden telah setidaknya mengkonsumsi 100 batang rokok selama masa hidupnya. Tabel 4.7 berisi rincian value dari variable Smoke.

Tabel 4.7 Variable Smoke

Label: Smoke at Least 100 Cigarettes
Section Name: Tobacco Use
Section Number: 9
Question Number: 1
Column: 193
Type of Variable: Num
SAS Variable Name: SMOKE100
Question Prologue:
Question: Have you smoked at least 100 cigarettes in your entire life? (Note: 5 packs = 100 cigarettes)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	204,239	43.43	40.14
2	No – Go to Section 09.05 USENOW3	262,670	55.33	58.33
7	Don't know/Not Sure – Go to Section 09.05 USENOW3	2,007	0.43	0.37

9	Refused – Go to Section 09.05 USENOW3	1,325	0.28	1.15
Blank	Not asked or Missing	16,062		

8. *OVERWEIGHT* – berisi informasi mengenai berat badan responden dengan nilai BMI paling kecil 25. Tabel 4.8menunjukkan rincian value dari variable Overweight.

Tabel 4.8 Variable Overweight

Label: Computed body mass index categories Section Name: Calculated Variables Section Number: 8 Question Number: 19 Column: 1996 Type of Variable: Num SAS Variable Name: _BMI5CAT Question Prologue: Question: Four-categories of Body Mass Index (BMI)				
Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Underweight Notes: _BMI5 < 1850 (_BMI5 has 2 implied decimal places)	7,530	1.69	1.98
2	Normal Weight Notes: 1850 <= _BMI5 < 2500	142,110	31.81	33.20

3	Overweight Notes: 2500 <= _BMI5 < 3000	161,282	36.11	35.24
4	Obese Notes: 3000 <= _BMI5 < 9999	135,765	30.39	29.58
Blank	Don't know/Refused/Missing Notes: _BMI5 = 9999	39,616		

9. *OBESE* – menunjukkan informasi berat badan responden dengan nilai BMI paling kecil 30. Tabel 4.9berisi rincian value variable Obese.

Tabel 4.9 Variable Obesity

Label: Overweight or obese calculated variable
Section Name: Calculated Variables
Section Number: 8
Question Number: 20
Column: 1997
Type of Variable: Num
SAS Variable Name: _RFBMI5
Question Prologue:
Question: Adults who have a body mass index greater than 25.00 (Overweight or Obese)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	No Notes: 1200 <= _BMI5 < 2500 (_BMI5 has 2 implied	149,640	30.77	32.01

	decimal places)			
2	Yes Notes: 2500 <= _BMI5 < 9999	297,047	61.08	58.97
9	Don't know/Refused /Missing Notes: _BMI5 = 9999	39,616	8.15	9.02

10. *PHYSICAL ACTIVITY* – berisi informasi apakah responden melakukan aktifitas fisik di sebulan terakhir selain yang berhubungan dengan pekerjaan atau aktivitas rutin responden. Tabel 4.10 menunjukkan rincian value dari variable Physical Activity.

Tabel 4.10 Variable Physical Activity

Label: Overweight or obese calculated variable Section Name: Calculated Variables Section Number: 8 Question Number: 20 Column: 1997 Type of Variable: Num SAS Variable Name: _RFBMI5 Question Prologue: Question: Adults who have a body mass index greater than 25.00 (Overweight or Obese)				
Value	Value Label	Frequency	Percentage	Weighted Percentage
1	No Notes: 1200 <= _BMI5 < 2500 (_BMI5 has 2 implied decimal	149,640	30.77	32.01

	places)			
2	Yes Notes: 2500 <= _BMI5 < 9999	297,047	61.08	58.97
9	Don't know/Refused /Missing Notes: _BMI5 = 9999	39,616	8.15	9.02

11. *DIABETES* – berisi informasi apakah responden pernah didiagnosa diabetes oleh tenaga kesehatan (dokter). Tabel 4.11 berisi rincian value dari variable Diabetes.

Tabel 4.11 Variable Diabetes

Label: (Ever told) you have diabetes Section Name: Chronic Health Conditions Section Number: 6 Question Number: 12 Column: 115 Type of Variable: Num SAS Variable Name: DIABETE3 Question Prologue: Question: (Ever told) you have diabetes (If 'Yes' and respondent is female, ask "was this only when you were pregnant?'. If Respondent says pre-diabetes or borderline diabetes, use response code 4.)				
Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Yes	66,053	13.58	10.80
2	Yes, but female told only during pregnancy – Go to Section 07.01	3,644	0.75	0.91

	LASTDEN3			
3	No – Go to Section 07.01	406,884	83.67	86.27
4	No, pre-diabetes or borderline diabetes – Go to Section 07.01 LASTDEN3	8,858	1.82	1.82
7	Don’t know/Not Sure – Go to Section 07.01 LASTDEN3	626	0.13	0.17
9	Refused – Go to Section 07.01 LASTDEN3	235	0.05	0.04
Blank	Not asked or Missing	3		

4.3 Persiapan Data

4.3.1 Pemilihan Data

Setelah pemilihan atribut variable yang diperlukan, pemilihan data selanjutnya didasarkan atas lengkap atau tidaknya data yang diperoleh, dalam hal ini berhubungan dengan variable Disposition yakni dipilih data dengan value 1100 yang berarti bahwa wawancara dilakukan secara lengkap tanpa terputus di tengah-tengah sehingga kemungkinan besar menghasilkan data yang cukup lengkap. Dalam proses selanjutnya, variable Disposition tidak digunakan dalam proses analisis karena hanya digunakan sebagai indikator awal layak atau tidaknya data untuk digunakan. Selain itu untuk variable Diabetes juga akan dilakukan pemilihan data yang bernilai 1 (*Yes*) dan 3 (*No*).

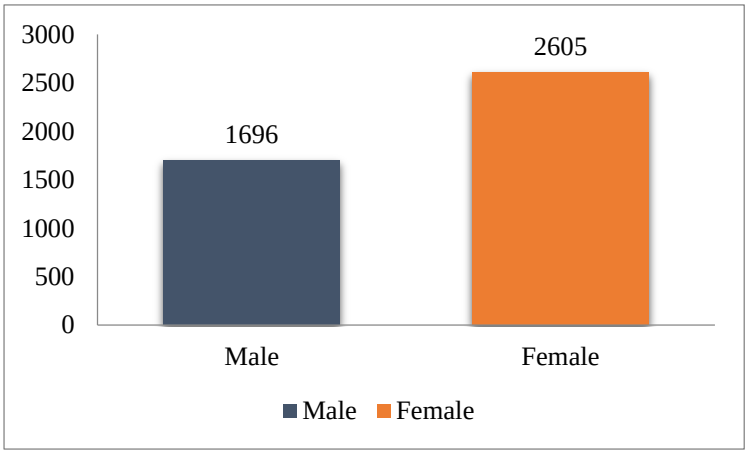
4.3.2 *Missing Value Treatment*

Proses *treatment missing value* adalah dengan menghapus seluruh baris yang terdapat *missing value.* Sampai proses ini, maka dihasilkan dataset sebanyak 4301 baris dengan 10 atribut variable yang akan digunakan pada proses selanjutnya yakni implementasi.

4.3.3 Statistik Data

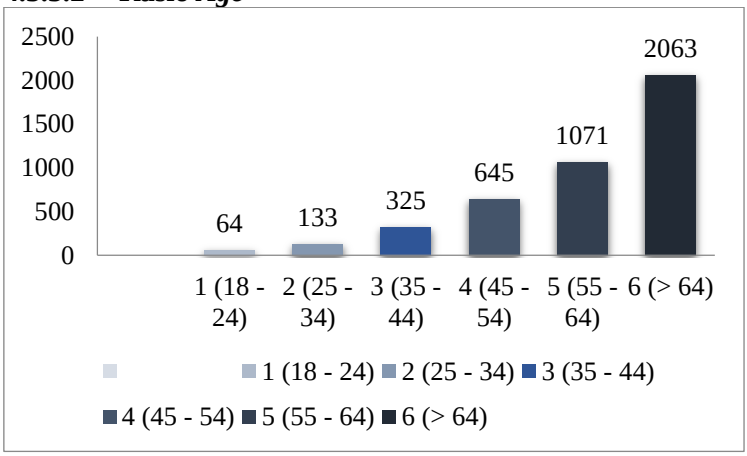
Setelah dilakukan praproses pemilihan data, maka berikut ini adalah statistik data-data yang akan digunakan.

4.3.3.1 Rasio *Gender*



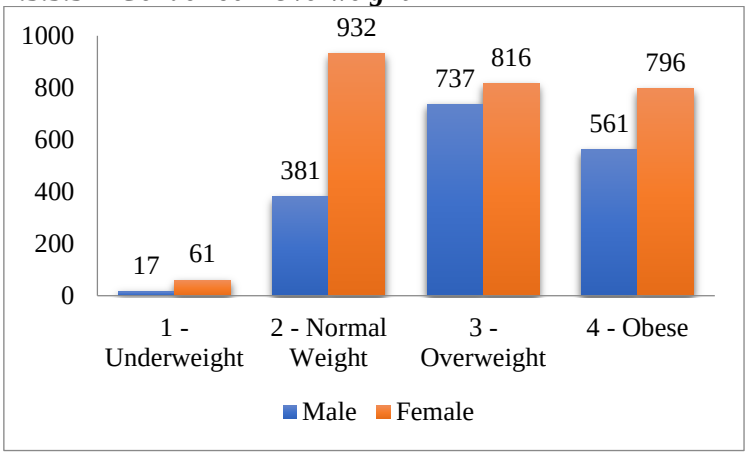
Gambar 4.3 Rasio Jenis Kelamin Responder

4.3.3.2 Rasio *Age*



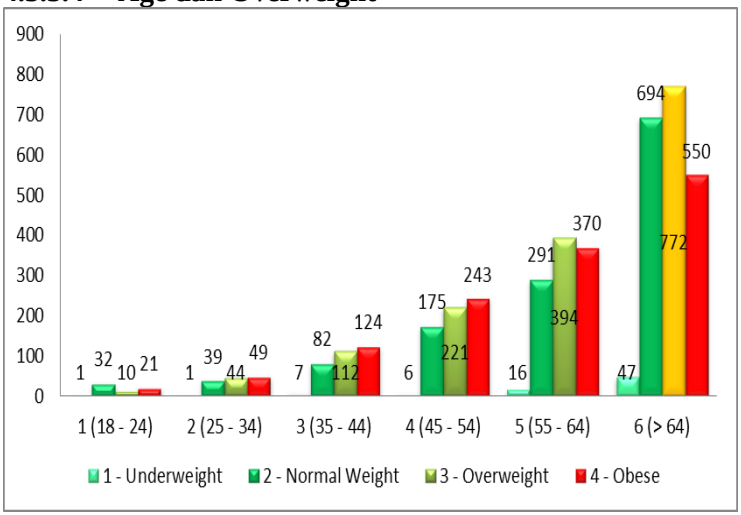
Gambar 4.4 Rasio Umur Responden

4.3.3.3 *Gender* dan *Overweight*



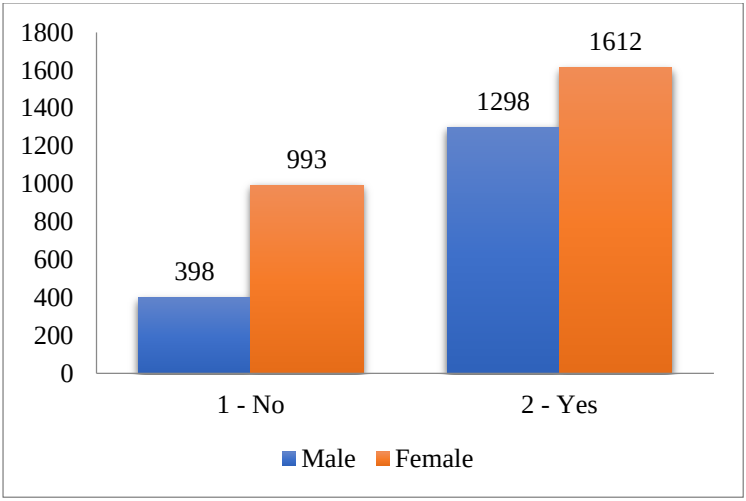
Gambar 4.5 Rasio Overweight Berdasarkan Jenis Kelamin

4.3.3.4 Age dan Overweight



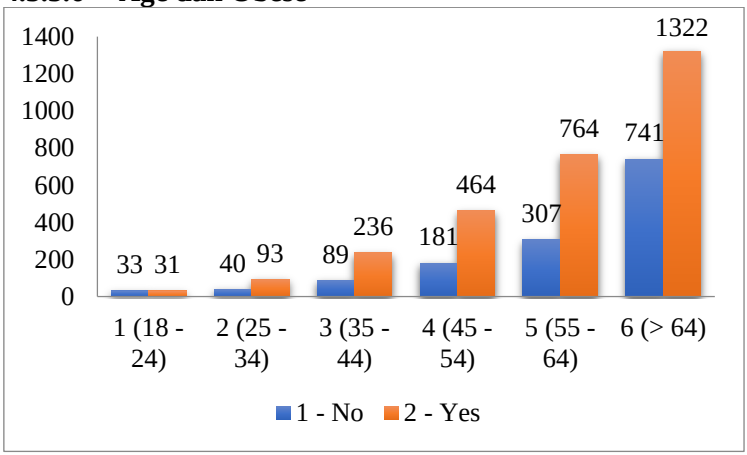
Gambar 4.6 Rasio Overweight Berdasarkan Usia

4.3.3.5 Gender dan Obese



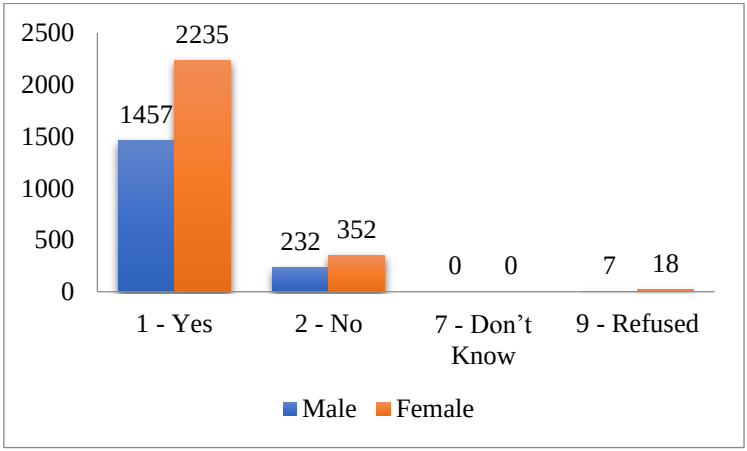
Gambar 4.7 Rasio Obese Berdasarkan Jenis Kelamin

4.3.3.6 Age dan Obese



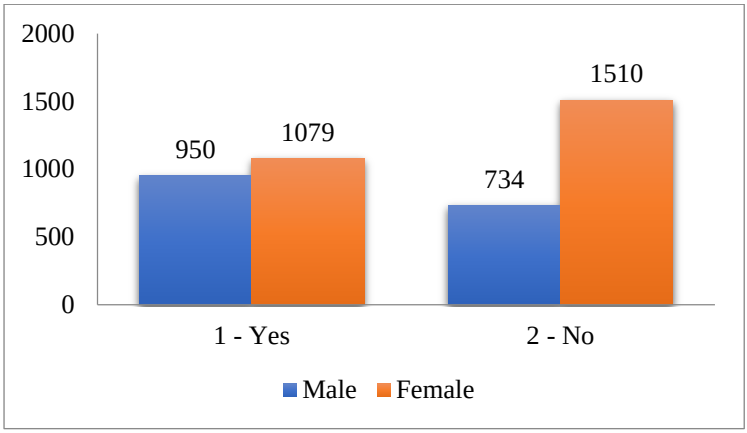
Gambar 4.8 Rasio Obese Berdasarkan Usia

4.3.3.7 Gender dan Smoke



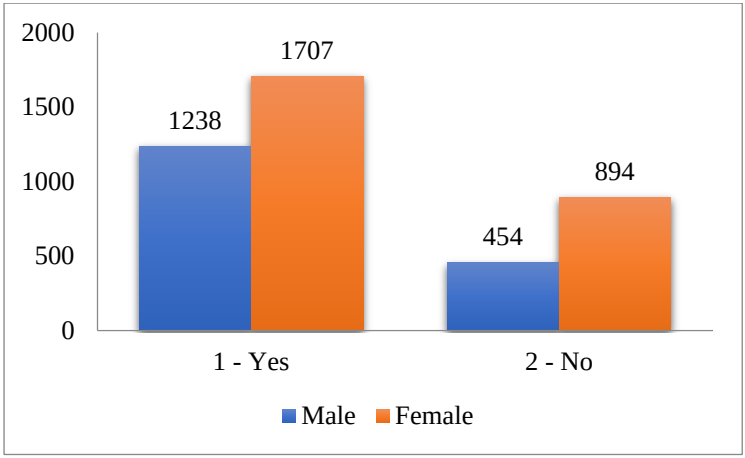
Gambar 4.9 Rasio Smoke Berdasarkan Jenis Kelamin

4.3.3.8 Gender dan Alcohol



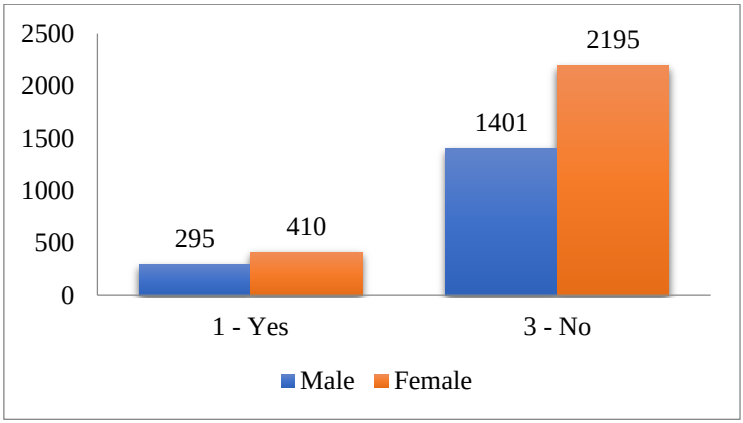
Gambar 4.10 Rasio Alcohol Berdasarkan Jenis Kelamin

4.3.3.9 Gender dan Physical Activity



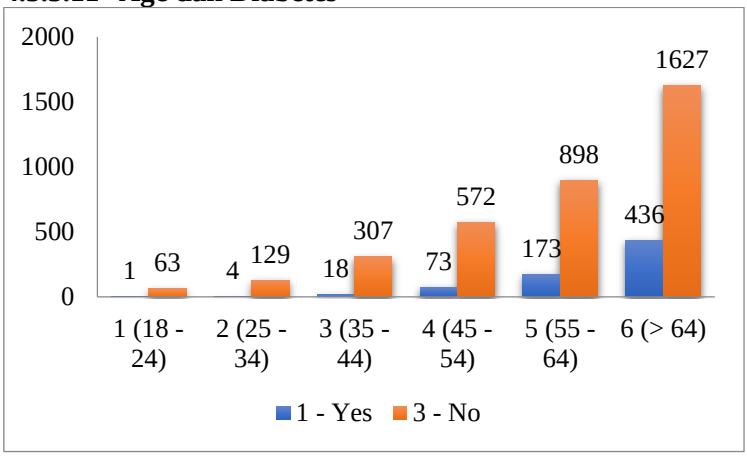
Gambar 4.11 Rasio Physical Activity Berdasarkan Jenis Kelamin

4.3.3.10 Gender dan Diabetes



Gambar 4.12 Rasio Diabetes Berdasarkan Jenis Kelamin

4.3.3.11 Age dan Diabetes



Gambar 4.13 Rasio Diabetes Berdasarkan Usia

4.3.4 Proses Prediksi SVM

Supervised learning antara lain terdiri dari proses mempelajari hubungan antara dua set data: yakni data yang diobservasi (x) dan variable eksternal (y) yang dicoba untuk diprediksi, dimana biasanya disebut "target" atau "label". Seringkali nilai y adalah *array* satu dimensi dari seluruh panjang n_*samples.* Seluruh dari estimator yang telah dipelajari akan mengimplementasikan *method fit (x,y)* untuk menyesuaikan ke dalam model dan *method predict (x)*, dimana nilai observasi dari x yang tidak berlabel akan dikembalikan ke pada nilai label prediksi y.

4.3.4.1 Training Set dan Testing Set

Data set akan dibagi menjadi *training set* dan *testing set* dimana pada penelitian ini adalah digunakan *classifier K-fold Cross Validation.*

4.3.4.2 Parameter Kernel

Class pada dataset yang ada kemungkinan tidak selalu terpisah secara linier dalam ruang fitur, sehingga dibutuhkan solusi untuk membangun ruang keputusan yang tidak linier atauun bahkan bisa bersifat polinominal. Oleh karena itu digunakan trik kernel yang paling tidak bisa memungkinkan keputusan prediksi yang lebih baik dengan melakukan obesrvasi penggunaan kernel yang lebih beragam. Diantara kernel yang akan digunakan adalah *Linear kernel* dan *Radial Basis Function (RBF) kernel*.

4.3.4.3 Optimasi Grid Search

Optimasi Grid Search adalah proses pencarian parameter terbaik dengan menggunakan system yang terkomputerisasi dan sistematis. Pencarian parameter terbaik berdasarkan kernel yang dipilih seperti *Linear* ataupun RBF dilakukan untuk mendapatkan performa yang terbaik yang bisa dicapai dan seoptimal mungkin.

4.3.4.4 Resample

Jika nantinya hasil prediksi menunjukkan performa yang kurang baik, kemungkinan disebabkan oleh data yang tidak

seimbang (*imbalanced data*). Oleh karena itu perlu dilakukan *resample* data untuk mendapatkan hasil prediksi yang optimal. *Resample* dapat dilakukan dengan *upsampling* atau *downsampling* ataupun bisa dilakukan keduanya jika dibutuhkan.

Halaman ini sengaja dikosongkan.

BAB VI
IMPLEMENTASI

Bab ini berisi tentang proses implementasi dalam pembuatan model prediksi yang dijalankan dengan algoritma SVM. Implementasi dioperasikan menggunakan pemograman bahasa Python pada program Anaconda menggunakan Jupyter Notebook.

5.1 Impor Library

Berikut ini adalah library yang perlu diimport untuk bisa menjalankan tahap prediksi selanjutnya.

```
# Library

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from statistics import mean

from sklearn import svm

from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from sklearn.metrics import mean_absolute_error
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.metrics import precision_recall_fscore_support
from sklearn.metrics import confusion_matrix
```

Kode 5.1 Library pada SVM

Berikut ini perjelasan dari *module library* yang digunakan:

Tabel 5.1 Module Library dan Fungsinya

Module	***Function***	***Utility***
sklearn		Singkatan dari *Scikit Learn* yang merupakan *library* pembelajaran mesin *open source* untuk bahasa pemograman *Python* yang efisien untuk *data mining* dan *data analysis*.
numpy		Menyediakan objek matematika yang mempermudah perhitungan matermatika. Objek yang disediakan ialah *array* dalam bentuk *matrix*.
matplotlib		Berguna untuk memvisualisasikan hasil perhitungan dengan berbagai macam grafik.
pandas		Digunakan untuk menyediakan struktur data dan analisis data yang mudah digunakan dan memiliki kinerja tinggi untuk bahasa pemograman *python. Pandas* memiliki struktur data yang diperlukan untuk membersihkan data mentah ke dalam sebuah bentuk data yang cocok

		untuk analisis.
statistics	*mean*	Menghitung nilai rata-rata.
sklearn.metrics	*classification_report*	Membuat laporan hasil klasifikasi
	accuracy_score	Menghitung nilai akurasi dari klasifikasi
	mean_absolute_error	Menghitung nilai *error* dari klasifikasi
	precision_recal_fscore_support	Menghitung nilai *precision, recall, fscore,* dan *support vector*
	confusion_matrix	Menghitung *confusion matrix* dari hasil klasifikasi
sklearn. model_ selections	*train_test_split*	Memmisahkan data yang berupa *array matrix* ke dalam subset *train set* dan *test set*
	KFold	Menyediakan indeks *train* atau *test* untuk membagi data ke dalam beberapa *fold*

5.2 Memuat data

Untuk dapat membaca data, maka dataset dengan format data *.csv* yang telah disiapkan harus dipanggil terlebih dahulu dan untuk melakukan pengecekan bahwa data yang diload sudah sesuai, lakukan perintah tampilkan 10 data teratas.

```
data = pd.read_csv('fun5.csv')

# cek data 10 teratas
data.head(10)
```

Kode 5.2. Load data

	STATE	SEX (SEX)	AGE (_AGE_G)	RACE (_RACE)	OVERWEIGHT	OBESE	SMOKE (SMOKE100)	DRINK	EXERANY2	DIABETES (DIABETE3)
0	1	1	3	1	2	1	1	2	1	3
1	1	2	5	1	3	2	1	2	1	3
2	1	2	6	1	2	1	2	1	1	3
3	1	1	6	1	3	2	1	2	1	1
4	1	1	1	1	2	1	1	2	1	3
5	1	2	6	1	2	1	1	1	1	3
6	1	2	6	1	4	2	1	2	1	1
7	1	2	3	1	4	2	1	2	1	3
8	1	1	5	1	4	2	1	1	1	3
9	1	2	6	1	2	1	1	2	1	3

Kode 5.3. Preview load data

5.3 Convert Data to Matrix

Data awal yang berekstensi csv terlebih dahulu dikonversi ke dalam bentuk matrix untuk bisa menjalankan fungsi *pandas*.

```
# Convert to matrix
data_matrix = data.as_matrix()
```

Kode 5.4. Konversi ke dalam bentuk matrix

5.4 Define Train and Test Data

Terdapat 10 atribut yang akan dijalankan pada proses ini, dimana array 0 – 8 menjadi data training dan array ke 9 akan menjadi data target.

```
data_train = np.array(data_matrix[:, 0:9])
data_target = np.array(data_matrix[:,9])

print(data_train)
print(data_target)
```

Kode 5.5. Pendefinisian data train dan data test

5.5 Running The Model

5.5.1 Cross Validation

Cross validation dilakukan untuk menghasilkan hasil prediksi dengan nilai akurasi yang tinggi dengan *fold* sebanyak 10 kali.

```
kf = KFold(n_splits=10)
```

Kode 5.6. Cross Validation

5.5.2 Menyimpan label aktual dan predicted dari seluruh Fold

Setelah dijalankannya perhitungan *cross validation*, maka akan dihasilkan serangian nilai tertentu dimana terapat dua label dengan satu label dengan nilai aktual dan label dengan nilai yang diprediksi.

```
final_test_labels = []
final_prediction = []
```

Kode 5.7. Penyimpanan nilai tabel aktual dan prediksi

5.5.3 Menyimpan seluruh hasil pengukuran dari setiap Fold

Dari masing-masing fold yang dilakukan, setiap foldnya akan menghasilkan perhitungan nilai mterik presisi, recall dan fscore.

```
precision_val = []
recall_val = []
fscore_val = []
```

Kode 5.8. Penyimpana perhitungan presisi, recall, dan fscore dari setiap fold

5.6 Define Parameter

Terdapat 4 tipe metode kalsfikasi dalamSVM, yakni Linier, RBF kernel, LinearSVC, dan polyniminal. Gambar dibawah ini menunjukkan perintah operasi menggunal kernel linear dengan nilai default C sebesar 1. Nilai C dan Gamma akan terus mengalami perubahan dalam masa percobaan untuk menghasilkan nilai akurasi yang paling baik.

```
# Tuning Parameter
C = 1
gamma = 'auto'
kernel = 'linier'
```

Kode 5.9. Tuning parameter

5.7 Implementasi *Training*

Perintah training dimana x mengindikasikan set untuk training dan y untuk destinasi hasil prediksi training.

```
for train_index, test_index in kf.split(data_train):

    # menentukan indeks yang dijadikan data training dan data testing
    # X adalah data atribut
    # y adalah data label

    X_trainSet, X_testSet = data_train[train_index], data_train[test_index]
    y_trainSet, y_testSet = data_target[train_index], data_target[test_index]

    # Start Train
    classifier = svm.SVC(kernel=kernel, C=C, gamma=gamma).fit(X_trainSet,y_trainSet)
```

Kode 5.10. Implementasi training

5.8 Implementasi *Testing*

Setelah dilakukan training, maka testing akan menjalankan perintah perhitungan performa prediksi berisi akurasi, presisi, recall, dan fscore.

```
# Start Testing
prediction = classifier.predict(X_testSet)

# Start perhitungan performa
precision, recall, fscore, Null_Value = precision_recall_fscore_support(y_testSet, prediction, average='macro')

# Print performance setiap fold
print(precision_recall_fscore_support(y_testSet, prediction, average='weighted'))

# Nilai hasil pengukuran ditambahkan ke array
precision_val.append(precision)
recall_val.append(recall)
fscore_val.append(fscore)

# menambahkan label per fold ke array
final_test_labels.extend(y_testSet)
final_prediction.extend(prediction)
```

Kode 5.11. Impelementasi testing

5.9 Performance

Setelah dihasilkan nilai hasil akurasi, presisi, dan recall pada testing dengan perintah kode seperti dibawah ini, maka akan didapatkan nilai rata-rata keseluruhannya dari ketiga hasil tersebut.

```
print("avg recall :", mean(recall_val))
print("avg presisi :", mean(precision_val))
print("akurasi :", accuracy_score(final_test_labels, final_prediction))
```

Kode 5.12. Perhitungan nilai rata-rata presisi, recall, dan fscore

```
avg recall : 0.23
avg presisi : 0.1862278570705537
akurasi : 0.8095452498874381
```

Kode 5.13. Hasil rata-rata perhitungan metrik

Sedangkan berikut ini adalah perintah dan hasil dari classification report.

```
print(classification_report(final_test_labels, final_prediction))
```

Kode 5.14. Classification report

	precision	recall	f1-score	support
1	0.00	0.00	0.00	705
2	0.00	0.00	0.00	38
3	0.81	1.00	0.89	3596
4	0.00	0.00	0.00	99
7	0.00	0.00	0.00	4
avg / total	0.66	0.81	0.72	4442

Kode 5.15. Hasil classification report

5.10 Grid Search

Untuk mengetahui dan menghasilkan nilai parameter terbaik, maka dilakukan optimasi menggunakan *grid serach*. Langkah pertama yang dilakukan adalah melakukan impor library yakni GridSeacrhCV.

```
from sklearn.model_selection import GridSearchCV
```

Kode 5.16 Impor Library GridSerachCV

Data kali ini dibagi menjadi 2 bagian seimbang antara data set x dan y untuk masing-masing kebutuhan *training* dan *test*.

```
# Split the dataset in two equal parts
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.5, random_state=0)
```

Kode 5.17 Pembagian data set ke dalam 2 bagian sama

Selanjutnya dilakukan penentuan rentang parameter yang ingin diuji. Beberapa nilai C dimasukkan yakni 1, 10, 100, dan 1000 untuk kedua fungsi kernel Linear maupun RBF, sedangkan nilai Gamma akan dicari diantara rentang 0.001 hingga 100.

```
# Set the parameters by cross-validation 'gamma': [1e-3, 1e-4]
tuned_parameters = [{'kernel': ['rbf'], 'gamma': [100, 10, 1, 1e-2, 1e-3, 1e-4],
                     'C': [1, 10, 100, 1000]},
                    {'kernel': ['linear'], 'C': [1e-1, 1, 10, 100, 1000]}]
```

Kode 5.18 Menentukan rentang parameter C dan Gamma

Selanjutnya adalah proses pencarian parameter terbaik menggunakan perintah seperti dibawah ini:

```
for score in scores:
    print("# Tuning hyper-parameters for %s" % score)
    print()

    clf = GridSearchCV(SVC(), tuned_parameters, cv=10,
                       scoring='%s_macro' % score)
    clf.fit(X_train, y_train)

    print("Best parameters set found on development set:")
    print()
    print(clf.best_params_)
    print()
    print("Grid scores on development set:")
    print()
    means = clf.cv_results_['mean_test_score']
    stds = clf.cv_results_['std_test_score']
    for mean, std, params in zip(means, stds, clf.cv_results_['params']):
        print("%0.3f (+/-%0.03f) for %r"
              % (mean, std * 2, params))
    print()

    print("Detailed classification report:")
    print()
    print("The model is trained on the full development set.")
    print("The scores are computed on the full evaluation set.")
    print()
    y_true, y_pred = y_test, clf.predict(X_test)
    print(classification_report(y_true, y_pred))
    print()
```

Kode 5.19 Mencari Parameter Terbaik

Setelah proses selesai dijalankan, berikut ini adalah contoh hasil dari proses pencarian parameter terbaik (Tabel 5.2), dimana disebutkan bahwa nilai parameter terbaik adalah dengan C sebesar 1 dan Gamma sebesar 1.

Tabel 5.2 Hasil Penentuan Parameter Terbaik Menggunakan GridSearch

Best parameters set found on development set: {'C': 1, 'gamma': 1, 'kernel': 'rbf'}
Grid scores on development set: 0.551 (+/-0.149) for {'C': 1, 'gamma': 100, 'kernel': 'rbf'} 0.551 (+/-0.149) for {'C': 1, 'gamma': 10, 'kernel': 'rbf'} 0.570 (+/-0.208) for {'C': 1, 'gamma': 1, 'kernel': 'rbf'} 0.417 (+/-0.002) for {'C': 1, 'gamma': 0.01, 'kernel': 'rbf'}

0.417 (+/-0.002) for {'C': 1, 'gamma': 0.001, 'kernel': 'rbf'}
0.417 (+/-0.002) for {'C': 1, 'gamma': 0.0001, 'kernel': 'rbf'}
0.551 (+/-0.149) for {'C': 10, 'gamma': 100, 'kernel': 'rbf'}
0.551 (+/-0.149) for {'C': 10, 'gamma': 10, 'kernel': 'rbf'}
0.538 (+/-0.113) for {'C': 10, 'gamma': 1, 'kernel': 'rbf'}
0.442 (+/-0.151) for {'C': 10, 'gamma': 0.01, 'kernel': 'rbf'}
0.417 (+/-0.002) for {'C': 10, 'gamma': 0.001, 'kernel': 'rbf'}
0.417 (+/-0.002) for {'C': 10, 'gamma': 0.0001, 'kernel': 'rbf'}
0.551 (+/-0.149) for {'C': 100, 'gamma': 100, 'kernel': 'rbf'}
0.551 (+/-0.149) for {'C': 100, 'gamma': 10, 'kernel': 'rbf'}
0.538 (+/-0.113) for {'C': 100, 'gamma': 1, 'kernel': 'rbf'}
0.547 (+/-0.259) for {'C': 100, 'gamma': 0.01, 'kernel': 'rbf'}
0.417 (+/-0.002) for {'C': 100, 'gamma': 0.001, 'kernel': 'rbf'}
0.417 (+/-0.002) for {'C': 100, 'gamma': 0.0001, 'kernel': 'rbf'}
0.551 (+/-0.149) for {'C': 1000, 'gamma': 100, 'kernel': 'rbf'}
0.551 (+/-0.149) for {'C': 1000, 'gamma': 10, 'kernel': 'rbf'}
0.538 (+/-0.113) for {'C': 1000, 'gamma': 1, 'kernel': 'rbf'}
0.559 (+/-0.195) for {'C': 1000, 'gamma': 0.01, 'kernel': 'rbf'}
0.417 (+/-0.002) for {'C': 1000, 'gamma': 0.001, 'kernel': 'rbf'}
0.417 (+/-0.002) for {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
0.417 (+/-0.002) for {'C': 0.1, 'kernel': 'linear'}
0.417 (+/-0.002) for {'C': 1, 'kernel': 'linear'}
0.417 (+/-0.002) for {'C': 10, 'kernel': 'linear'}
0.417 (+/-0.002) for {'C': 100, 'kernel': 'linear'}
0.417 (+/-0.002) for {'C': 1000, 'kernel': 'linear'}

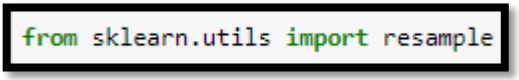
Detailed classification report:

The model is trained on the full development set.
The scores are computed on the full evaluation set.

	precision	recall	f1-score	support
1	0.43	0.05	0.08	348
3	0.84	0.99	0.91	1803
avg / total	0.78	0.84	0.78	2151

5.11 Resample

Untuk *data imbalance*, maka perlu dilakukan resample. Langkah awal yakni dengan melakukan impor library seperti berikut.



Kode 5.20 Impor Library Resample

Langkah selanjutnya adalah loada data dengan perintah seperti dibawah ini. Dalam label Diabetes, data major adalah yang bernilai 3 sedangkan data minor bernilai 1. Maka dilakukan Downsampling data major sebanyak jumlah data minor yang ada yakni 705 bari data.

```
data = pd.read_csv('fun7.csv')

#downsampling

data['DIABETES'].value_counts()

data_majority = data[data.DIABETES==3]

data_minority = data[data.DIABETES==1]

data_majority_downsampled = resample(data_majority,
replace=False, n_samples=705, random_state=123)

data_sampled = pd.concat([data_majority_downsampled,
data_minority])

data_sampled.DIABETES.value_counts()

data_sampled.head(10)
```

Kode 5.21 Downsampling Data

Langkah selanjutnya adalah sama dengan ketika melakukan prediksi pada awal percobaan.

BAB VI
HASIL DAN PEMBAHASAN

Pada bab ini akan dijelaskan hasil serta analisis terhadap hasil yang diperoleh dari proses implementasi yang telah dibahas pada bab sebelumnya.

6.1 Prediksi Menggunakan SVM Kernel Linear

Berikut ini adalah hasil nilai akurasi prediksi SVM menggunakan kernel linear beserta *confusion matrix* dengan niali C antara 0.001 hingga 1000.

Tabel 6.1 Hasil Prediksi Menggunakan Kernel Linear

Kernel	C	Accuracy	Confusion Matrix
Linear	0.001	0.8360	[[0 705] [0 3596]]
	0.01	0.8360	[[0 705] [0 3596]]
	0.1	0,8360	[[0 705] [0 3596]]
	1	0,8360	[[0 705] [0 3596]]
	10	0,8360	[[0 705] [0 3596]]
	100	0,8360	[[0 705] [0 3596]]
	1000	0,8360	[[0 705] [0 3596]]

6.1.1 Nilai Presisi, Recall, dan Fscore

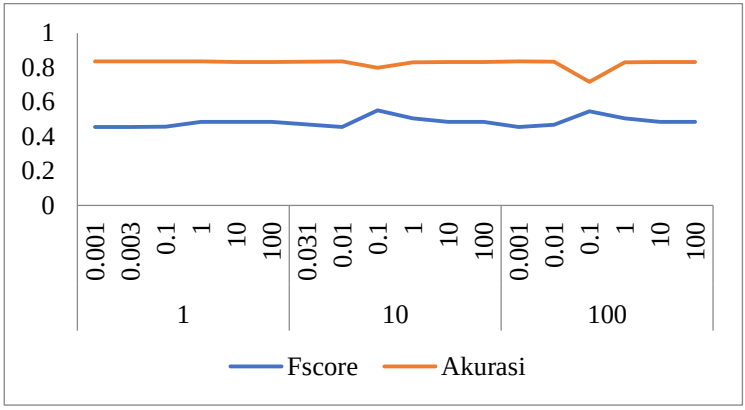
Berikut ini adalah nilai presisi, recall dan fscore dari hasil prediksi SVM menggunakan metode linear dengan hasil nilai akurasi terbaik.

Tabel 6.2 Nilai Presisi, Recall dan Fscore dari Akurasi Tertinggi Menggunakan Kernel Linear

- avg recall : 0.5
- avg presisi : 0.4180453785140023
- avg fscore : 0.455339481073065
- akurasi : 0.8360846314810509

6.2 Prediksi Menggunakan SVM Kernel Rbf

Berikut ini (Gambar 6.1) adalah grafik yang menunjukkan hubungan pengaruh antara nilai C dan Gamma terhadap nilai Fscore dan Akurasi. Ditunjukkan bahwa nilai akurasi terbaik sebesar 83.60% adalah dengan nilai C 1 dan Gamma 0.001.



Gambar 6.1 Grafik Hubungan Nilai C dan Gamma Terhadap Nilai Fscore dan Akurasi

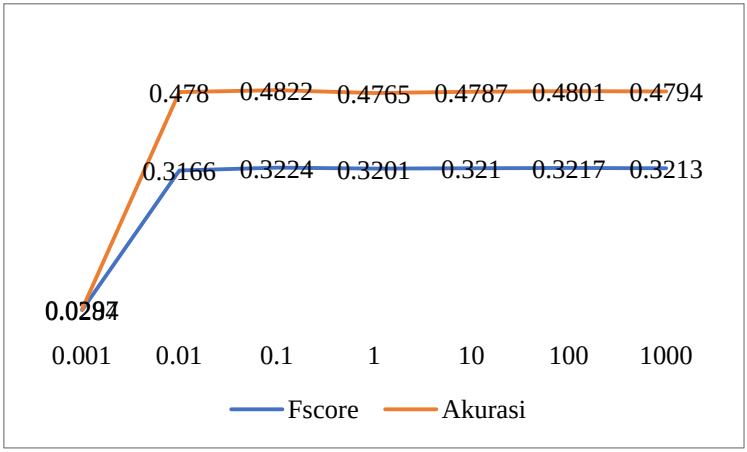
Namun jika memperhitungakan nilai *confusion matrix*, hasil paling baik ditunjukkan pada C sebesar 100 dan Gamma sebesar 0.1, namun menghasilkan akurasi yang tidak begitu

tinggi yakni sebesar 71.72% saja. Untuk hasil perhitungan beserta nilai *confusion matrix* bisa dilihat pada Lampiran A.

6.3 Downsampling

Berikut ini adalah hasil nilai akurasi prediksi SVM baik menggunakan kernel Linear (Gambar 6.2) ataupun RBF (Gambar 6.3) ketika diterapkan downsampling. Ditunjukkan bahwa nilai akurasi yang dihasilkan sangat rendah yakni hanya berkisar antara 2.97% dengan hasil confusion matrix yang tidak optimal. Nilai akurasi terbaik didapatkan adalah sebesar 4.82% untuk C 0.1 dengan hasil *confusion matrix* yang lumayan baik seperti berikut untuk kernel Linear. Untuk melihat nilai hasil lebih detil, lihat Lampiran B.

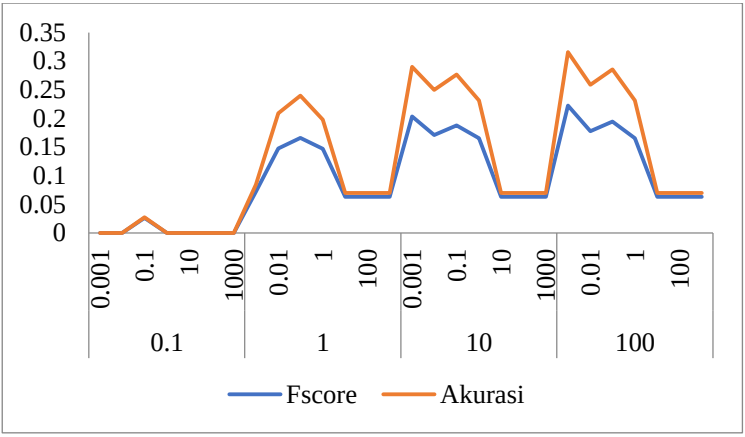
```
[[318 387]
 [343 362]]
```



Gambar 6.2 Grafik Hubungan Nilai C Terhadap Nilai Fscore dan Akurasi Kernel Linear Setelah Dilakukan Dowsampling

Sedangkan nilai akurasi terbaik untuk kernel RBF yang didapatkan adalah sebesar 3.15% untuk C 100 dan Gamma 0.001 dengan hasil *confusion matrix* seperti dibawah ini. Untuk melihat nilai hasil lebih detil, lihat Lampiran B.

```
[[ 96 609]
 [356 349]]
```



Gambar 6.3 Grafik Hubungan Nilai C dan Gamma Terhadap Nilai Fscore dan Akurasi Kernel RBF Setelah Dilakukan Dowsampling

6.4 Hasil Pembahasan

Pada model prediksi menggunakan metode Linear, didapatkan akurasi tertinggi yang sama sebesar 83.60% untuk nilai C = 0.001 hingga 1000, dengan confusion matrix seperti berikut. Hal ini menunjukkan bahwa model Linear tidak cocok digunakan untuk jenis tipe data ini karena hanya mampu melakukan prediksi terhadap nilai prediksi false.

```
[[   0  705]
 [   0 3596]]
```

Sedangkan pada model prediksi menggunakan metode RBF, didapatkan akurasi tertinggi yakni sebesar 71.72% untuk nilai C = 100 dan nilai Gamma = 0.1, dengan hasil confusion matrix seperti berikut. Dari hasil ini nilai aktual positif yang diprediksi benar sebanyak 243, dan nilai aktual salah yang diprediksi salah sebanyak 2842.

```
[[ 243  462]
 [ 754 2842]]
```

Proses lain yang dilakukan untuk meningkatkan performa pada imbalance data adalah dengan resample dimana dilakukan downsample pada data major sebesar jumlah data minor, namun hal ini juga tidak bisa membantu peningkatan performa melainkan menghasilkan nilai akurasi yang jauh dibawah hasil sebelumnya. Hal ini bisa membuktikan bahwa tidak semua jenis data ketika mengalami imbalance dapat langsung menghasilkan performa seperti yang diharapkan ketika menggunakan resample.

Jika dibandingkan dengan hasil penelitian sebelumnya yang menggunakan metode *Linear Regression,* dimana didapatkan hasil prediksi nilai benar yang tidak optimal seperti berikut (Gambar 6.4). Hal ini menunjukkan bahwa metode SVM juga tidak terlalu optimal ketika digunakan untuk memprediksi jenis data ini.

	Model–diabetes Yes---(0)	Model–diabetes No---(1)	Total
actual–diabetes yes---(0)	12	8805	8817
actual–diabetes no---(1)	21	78654	78675
Total	33	87459	87492

Gambar 6.4. Hasil penelitian sebelumnya

BAB VII
KESIMPULAN DAN SARAN

Pada bab ini dibahas mengenai kesimpulan dari semua proses yang telah dilakukan dan saran yang dapat diberikan untuk pengembangan yang lebih baik.

7.1 Kesimpulan

1. Performa model prediksi SVM terhadap training data menghasilkan akurasi 83.60% dengan penggunaan metode SVM Linier.Tidak adanya perubahan nilai akurasi pada penerapan metode SVM Linier yakni dengan mengubah nilai C menjadi 0,001 hingga 1000. Nilai presisi, recall, dan fscore berturut-turut adalah 50%; 41.80%; dan 45.53%.
2. Performa model prediksi SVM terhadap training data menghasilkan akurasi tertinggi sebesar 71.72% dengan nilai C = 100 dan niai Gamma = 0.1. Dengan nilai recall, presisi, fscore berturut-turut adalah 56.59%, 56.22%, dan 54.65%.
3. GridSearch untuk tipe data ini tidak berjalan secara optimal, karena hasil yang didapatkan ketika diimplementasikan menghasilkan akurasi yang cukup baik namun menghasilkan confusion matrix yang tidak terlalu optimal.
4. Proses pemngkatan performa prediksi dengan cara melakukan downsampling pada data yang tidak seimbang ini, juga tidak menghasilkan peningkatan melainkan sebaliknya menurunkan performa nilai akurasi.

7.2 Saran

- Dari pengerjaan tugas akhir ini, terdapat beberapa saran guna pengembangan penelitian ke depan.

1. Pada penelitian kali ini, penulis hanya menggunakan 2 jenis metode kernel yakni kernel linear dan RBF. Dari hasil yang didapatkan seperti metode kernel tidak begitu cocok dengan karakter data yang digunakan, oleh karena itu perlu dilakukan percobaan dengan menggunakan metode kernel lainnya untuk menghasilkan performa terbaik.
2. Metode yang digunakan dalam penelitian ini hanyasalah satu metode yakni SVM, mungkin akan lebih baik jika penelitian selanjutnya menggunakan metode lain seperti Naïve Bayes, Decision Tree, atau Artificial Neual Network sehingga nantinya hasil yang didapatkan bisa menjadi perbandingan untuk mengetahui algoritma mana yang menghasilkan performa terbaik.
3. Untuk melakukan *handling* pada *imbalance data*, perlu dilakukan metode lain selain resample data selain downsample, seperti upsampling atau melakukan perbaikan pada model yang dibangun beserta *evaluation metrics* yang digunakan.
4. Pada penelitian ini, tidak dilakukan penggalian mengenai variable faktor risiko behavioral apa yang mempengeruhi secara signifikan pada kasus diabetes, oleh karena itu alangkah lebih baiknya jika penelitian selanjutnya mempertimbangkan untuk melakukan analisis variable yang berpengaruh signifikan.

DAFTAR PUSTAKA

[1] World Health Organization, December, 2017, <http://www.who.int/news-room/fact-sheets/detail/diabetes>

[2] InfoDatin Pusat Data dan Informasi Kementrian Kesehatan RI, November, 2014, <http://www.depkes.go.id/resources/download/pusdatin/infodatin/infodatin-diabetes.pdf>

[3] Behavioral Risk Factor Surveillance System, May, 2014, <https://www.cdc.gov/brfss/about/index.htm>

[4] U.S. Department of Health & Human Services, Centers for Disease Control and Prevention, December, 2017, <https://www.cdc.gov/brfss/annual_data/annual_2016.html>

[5] Jay Pedersen, Fangyao Liu, Fahad Alfarraj, and Harry Ngondo, "Examining Disease Risk Factors by Mining Publicly Available Inormation," *Procedia Computer Science 17 - Information Technology and Quantitative Management*, pp. 48 - 53, 2013.

[6] Ali H. Mohkad et al., "Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors, 2001," vol. 289, no. 1, pp. 76-79, January 2003.

[7] Vangie Beal. Webopedia. March, 2016, <https://www.webopedia.com/TERM/A/ASCII_file.html>

[8] Aditya Satrya Wibawa and Ayu Purwarianti, "Indonesian Named-entity Recognition for 15 Classes Using," in *5th Workshop on Spoken*

Language Technology for Under-resourced Languages, Yogyakarta, Indonesia, 2016.

[9] Indra Budi, Stéphane Bressan, Gatot Wahyudi, and Zainal A. Hasibuan, "Named Entity Recognition for the Indonesian Language: Combining Contextual, Morphological and Part-of-Speech Features into a Knowledge Engineering Approach," in *8th International Conference, DS*, Singapore, 2005.

[10] Mohammad B. Hossin and Md Nasir Sulaiman, "A Review On Evaluation Metrics For Data Classification Evaluations," *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol. 5, no. 2, March 2015.

LAMPIRAN A – HASIL PREDIKSI SVM MENGGUNAKAN METODE RBF

	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	1	0.001	avg recall : 0.5 avg presisi : 0.41804537851400203 avg fscore : 0.455339481073065 akurasi : 0.836084631481050 9	[[0 705] [0 3596]]
	1	0.003	avg recall : 0.5 avg presisi : 0.41804537851400203 avg fscore : 0.455339481073065 akurasi : 0.836084631481050 9	[[0 705] [0 3596]]

A-2

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	1	0.1	avg recall : 0.5005080259562842 avg presisi : 0.43477685646545 71 avg fscore : 0.4567537582332727 akurasi : 0.8358521274122297	[[1 704] [2 3594]]
	1	1	avg recall : 0.5150484174685663 avg presisi : 0.5278081122546747 avg fscore : 0.4856518464812703 akurasi : 0.8353871192745873	[[26 679] [29 3567]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	1	10	avg recall : 0.5131101032391873 avg presisi : 0.5192785586609396 avg fscore : 0.4852104319649429 akurasi : 0.8321320623110905	[[26 679] [43 3553]]
	1	100	avg recall : 0.5131101032391873 avg presisi : 0.5192785586609396 avg fscore : 0.4852104319649429 akurasi : 0.8321320623110905	[[26 679] [43 3553]]

A-4

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	10	0.031	avg recall : 0.5061157952213842 avg presisi : 0.4992121151000492 avg fscore : 0.4706389489358948 akurasi : 0.8337595907928389	[[12 693] [22 3574]]
	10	0.01	avg recall : 0.4995878975216475 6 avg presisi : 0.41799171125103246 avg fscore : 0.4551345981800113 akurasi : 0.8353871192745873	[[0 705] [3 3593]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	10	0.1	avg recall : 0.5527440005778090 avg presisi : 0.5935736892919797 avg fscore : 0.5517684704200819 akurasi : 0.7995814926761219	[[129 576] [286 3310]]
	10	1	avg recall : 0.5241091036457519 avg presisi : 0.5922929481115562 avg fscore : 0.5052265450005712 akurasi : 0.8305045338293420	[[46 659] [70 3526]]

A-6

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	10	10	avg recall : 0.5131101032391873 avg presisi : 0.5192785586609396 avg fscore : 0.4852104319649429 akurasi : 0.8321320623110905	[[26 679] [43 3553]]
	10	100	avg recall : 0.5131101032391873 avg presisi : 0.5192785586609396 avg fscore : 0.4852104319649429 akurasi : 0.8321320623110905	[[26 679] [43 3553]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	100	0.001	avg recall : 0.5 avg presisi : 0.41804537851400203 avg fscore : 0.455339481073065 akurasi : 0.8360846314810509	[[0 705] [0 3596]]
	100	0.01	avg recall : 0.5045630463662916 avg presisi : 0.55491745978009887 avg fscore : 0.468626448309819 akurasi : 0.8332945826551965	[[11 694] [23 3573]]

A-8

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	100	0.1	avg recall : 0.5659443651951283 avg presisi : 0.5622912886584367 avg fscore : 0.54646084181343 akurasi : 0.7172750523134155	[[243 462] [754 2842]]
	100	1	avg recall : 0.5241091036457519 avg presisi : 0.5922929481115562 avg fscore : 0.5052265450005712 akurasi : 0.830504533829342	[[46 659] [70 3526]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	100	10	avg recall : 0.5131101032391873 avg presisi : 0.5192785586609396 avg fscore : 0.4852104319649429 akurasi : 0.8321320623110905	[[26 679] [43 3553]]
	100	100	avg recall : 0.5131101032391873 avg presisi : 0.5192785586609396 avg fscore : 0.4852104319649429 akurasi : 0.8321320623110905	[[26 679] [43 3553]]

Halaman ini sengaja dikosongkan

LAMPIRAN B – HASIL DOWNSAMPLING

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
Linear	0.001		avg recall : 0.014893617021276596 avg presisi : 0.4 avg fscore : 0.028439404594531967 akurasi : 0.029787234042553193	[[6 699] [669 36]]
	0.01		avg recall : 0.2390070921985815 avg presisi : 0.5 avg fscore : 0.3166906713706881 akurasi : 0.4780141843971631 4	[[324 381] [355 350]]

B-2

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
Linear	0.1		avg recall : 0.2411347517730496 avg presisi : 0.5 avg fscore : 0.3224408939846934 akurasi : 0.48226950354609927	[[318 387] [343 362]]
	1		avg recall : 0.2382978723404255 avg presisi : 0.5 avg fscore : 0.3201130310330133 akurasi : 0.4765957446808511	[[314 391] [347 358]]
	10		avg recall : 0.2393617021276596 avg presisi : 0.5 avg fscore : 0.3210627245981 akurasi : 0.4787234042553192	[[316 389] [346 359]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
Linear	100		avg recall : 0.2400709219858156 2 avg presisi : 0.5 avg fscore : 0.3217105785669708 3 akurasi : 0.4801418439716312 3	[[317 388] [345 360]]
	1000		avg recall : 0.2397163120567376 avg presisi : 0.5 avg fscore : 0.3213633050336473 5 akurasi : 0.4794326241134752	[[317 388] [346 359]]

B-4

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	0.1	0.001	avg recall : 0.0 avg presisi : 0.0 avg fscore : 0.0 akurasi : 0.0	[[0 705] [705 0]]
		0.01	avg recall : 0.0 avg presisi : 0.0 avg fscore : 0.0 akurasi : 0.0	[[0 705] [705 0]]
		0.1	avg recall : 0.0138297872340425 54 avg presisi : 0.25 avg fscore : 0.0261294844027965 1 akurasi : 0.0276595744680851 05	[[0 705] [666 39]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF		1	avg recall : 0.0 avg presisi : 0.0 avg fscore : 0.0 akurasi : 0.0	[[0 705] [705 0]]
		10	avg recall : 0.0 avg presisi : 0.0 avg fscore : 0.0 akurasi : 0.0	[[0 705] [705 0]]
		100	avg recall : 0.0 avg presisi : 0.0 avg fscore : 0.0 akurasi : 0.0	[[0 705] [705 0]]
		1000	avg recall : 0.0 avg presisi : 0.0 avg fscore : 0.0 akurasi : 0.0	[[0 705] [705 0]]

B-6

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	1	0.001	avg recall : 0.042198581560283686 avg presisi : 0.3 avg fscore : 0.0723450720421850 4 akurasi : 0.08439716312056737	[[1 704] [587 118]]
	1	0.01	avg recall : 0.1046099290780141 8 avg presisi : 0.3 avg fscore : 0.1475882543740839 6 akurasi : 0.20921985815602837	[[1 704] [411 294]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	1	0.1	avg recall : 0.1198581560283687 avg presisi : 0.4 avg fscore : 0.1659280743963775 akurasi : 0.2397163120567376	[[12 693] [379 326]]
	1	1	avg recall : 0.0989361702127659 avg presisi : 0.4 avg fscore : 0.1472474171619131 akurasi : 0.19787234042553192	[[20 685] [446 259]]
	1	10	avg recall : 0.0351063829787234 06 avg presisi : 0.4 avg fscore : 0.0630168758992653 akurasi : 0.07021276595744681	[[12 693] [618 87]]

B-8

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	1	100	avg recall : 0.035106382978723406 avg presisi : 0.4 avg fscore : 0.0630168758992653 akurasi : 0.07021276595744681	[[12 693] [618 87]]
	1	1000	avg recall : 0.035106382978723406 avg presisi : 0.4 avg fscore : 0.0630168758992653 akurasi : 0.07021276595744681	[[12 693] [618 87]
	10	0.001	avg recall : 0.1450354609929078 avg presisi : 0.5 avg fscore : 0.2037074894666470 akurasi : 0.2900709219858156	[[64 641] [360 345]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	10	0.01	avg recall : 0.1248226950354609 avg presisi : 0.35 avg fscore : 0.1708941098978350 akurasi : 0.2496453900 7092199	[[13 692] [366 339]]
	10	0.1	avg recall : 0.1382978723404255 avg presisi : 0.45 avg fscore : 0.1879620016276754 akurasi : 0.2765957446808511	[[29 676] [344 361]]
	10	1	avg recall : 0.1156028368794326 avg presisi : 0.45 avg fscore : 0.1655039404700851 akurasi : 0.2312056737 5886525	[[24 681] [403 302]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	10	10	avg recall : 0.035106382978723406 avg presisi : 0.4 avg fscore : 0.0630168758992653 akurasi : 0.07021276595744681	[[12 693] [618 87]]
	10	100	avg recall : 0.035106382978723406 avg presisi : 0.4 avg fscore : 0.0630168758992653 akurasi : 0.07021276595744681	[[12 693] [618 87]]
	10	1000	avg recall : 0.035106382978723406 avg presisi : 0.4 avg fscore : 0.0630168758992653 akurasi : 0.07021276595744681	[[12 693] [618 87]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	100	0.001	avg recall : 0.1578014184397163 2 avg presisi : 0.5 avg fscore : 0.2222032480863923 3 akurasi : 0.31560283687943264	[[96 609] [356 349]]
	100	0.01	avg recall : 0.1294326241134751 7 avg presisi : 0.45 avg fscore : 0.1775368892445854 akurasi : 0.25886524822695034	[[20 685] [360 345]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	100	0.1	avg recall : 0.1429078014184397 avg presisi : 0.5 avg fscore : 0.1946886518329025 akurasi : 0.28581560283687946	[[37 668] [339 366]]
	100	1	avg recall : 0.1156028368794326 avg presisi : 0.45 avg fscore : 0.1655039404700851 akurasi : 0.23120567375886525	[[24 681] [403 302]]
	100	10	avg recall : 0.0351063829787234 06 avg presisi : 0.4 avg fscore : 0.0630168758992653 1 akurasi : 0.07021276595744681	[[12 693] [618 87]]

Kernel	C	Gamma	Recall, Presisi, Fscore, Akurasi	Confusion Matrix
RBF	100	100	avg recall : 0.035106382978723406 avg presisi : 0.4 avg fscore : 0.0630168758992653 1 akurasi : 0.07021276595744681	[[12 693] [618 87]]
	100	1000	avg recall : 0.035106382978723406 avg presisi : 0.4 avg fscore : 0.0630168758992653 1 akurasi : 0.07021276595744681	[[12 693] [618 87]]

Halaman ini sengaja dikosongkan

BIODATA PENULIS

Penulis bernama Annisa Nurlaily, lahir di Rembang, Jawa Tengah pada tanggal 20 Agustus 1993. Merupakan anak pertama dari dua bersaudara. Penulis telah menempuh pendidikan formal di SD Negeri Kutoharjo 2 Rembang, SMP Negeri 2 Rembang, dan SMA Negeri 1 Pati.

Pada tahun 2011, penulis melanjutkan studi ke jenjang pendidikan yang lebih tinggi di Institut Teknologi Sepuluh Nopember sebagai mahasiswa Departemen Sistem Informasi, Fakultas Teknologi Informasi dan Komunikasi. Penulis terdaftar sebagai mahasiswa dengan nomor induk (NRP) 5211100123. Apabila terdapat pertanyaan mengenai Tugas Akhir ini, penulis dapat dihubungi melalui e-mail annisanurlaily@gmail.com