TUGAS AKHIR – KS141501

PREDIKSI DIAGNOSA KANKER SERVIKS BERDASARKAN INFORMASI DEMOGRAFI, KEBIASAAN, DAN REKAM MEDIS MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE*

PREDICTION OF CERVICAL CANCER DIAGNOSIS BASED ON DEMOGRAPHIC INFORMATION, HABITS, AND MEDICAL RECORDS USING SUPPORT VECTOR MACHINE ALGORITHM

SITI HAWA AMINAH
NRP 0521 12 4000 0054

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

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

TUGAS AKHIR – KS141501

PREDIKSI DIAGNOSA KANKER SERVIKS BERDASARKAN INFORMASI DEMOGRAFI, KEBIASAAN, DAN REKAM MEDIS MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE*

SITI HAWA AMINAH
NRP 05211240000054

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

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

TUGAS AKHIR – KS141501

PREDICTION OF CERVICAL CANCER DIAGNOSIS BASED ON DEMOGRAPHIC INFORMATION, HABITS, AND MEDICAL RECORDS USING SUPPORT VECTOR MACHINE ALGORITHM

SITI HAWA AMINAH
NRP 05211240000054

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

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

LEMBAR PENGESAHAN

PREDIKSI DIAGNOSA KANKER SERVIKS BERDASARKAN INFORMASI DEMOGRAFI, KEBIASAAN, DAN REKAM MEDIS 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:

SITI HAWA AMINAH
NRP. 0521 12 4000 0054

Surabaya, 25 Juli 2018

KEPALA
DEPARTEMEN SISTEM INFORMASI

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

LEMBAR PERSETUJUAN

PREDIKSI DIAGNOSA KANKER SERVIKS BERDASARKAN INFORMASI DEMOGRAFI, KEBIASAAN, DAN REKAM MEDIS 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:

SITI HAWA AMINAH
NRP. 0521 12 4000 0054

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)**

Radityo Prasetianto W., S.Kom., M.Kom. **(Penguji II)**

PREDIKSI DIAGNOSA KANKER SERVIKS BERDASARKAN INFORMASI DEMOGRAFI, KEBIASAAN, DAN REKAM MEDIS MENGGUNAKAN ALGORITMA *SUPPORT VECTOR MACHINE*

Nama Mahasiswa	**: Siti Hawa Aminah**
NRP	**: 05211240000054**
Departemen	**: Sistem Informasi FTIK-ITS**
Dosen Pembimbing	**: Renny Pradina, S.T., M.T., SCJP**

ABSTRAK

Penyakit kanker serviks merupakan salah satu penyakit yang dianggap sebagai penyakit paling mematikan di seluruh dunia. Menurut International Agency for Research on Cancer 2012, kanker serviks menempati peringkat ketiga pada penyakit yang sering diderita oleh wanita di seluruh dunia, dan menempati peringkat kedua di Indonesia. Menurut para ahli kanker, kanker serviks adalah salah satu jenis kanker yang paling dapat dicegah dan paling dapat disembuhkan dari semua kasus kanker. Salah satu kegiatan deteksi dini kanker serviks yang paling umum di Indonesia adalah menggunakan metode pap smear. Namun tersedianya data histori rekam medis pasien tidak disertai dengan proses ekstraksi data menjadi sebuah informasi yang dapat berguna untuk keputusan klinis [1].

Konsep klasifikasi merupakan bagian dari teknik data mining yang memiliki pekerjaan utama melakukan analis prediksi. Telah terdapat banyak penelitian untuk melakukan prediksi dan klasifikasi pada kasus medis, seperti pada penelitian [2] *dan* [3] *Dua penelitian tersebut menghasilkan klasifikasi dengan tingkat akurasi yang cukup tinggi, dengan nilai akurasi > 80 %. Sehingga diperlukan lebih banyak penelitian untuk mendapatkan suatu model yang mampu mengklasifikasikan*

dengan tingkat kesalahan minimal dan model yang dihasilkan dapat digunakan untuk melakukan prediksi data berikutnya.

Penelitian ini menggunakan metode Cross Validation agar proses training lebih akurat dan menghasilkan prediksi yang lebih baik. Percobaan prediksi dilakukan dengan SVM kernel linear dan RBF. Atribut yang digunakan berjumlah 27 atribut, dengan 1 atribut target yaitu hasil tes biopsy pasien. Data yang digunakan berjumlah 668 data dan 200 data yang telah dilakukan resample data.

Hasil dari penelitian ini merupakan hasil prediksi yang dilakukan pada dataset dari pasien 'Hospital Universitario de Caracas' di Caracas, Venezuela, dengan menggunakan algoritma Support Vector Machine. Hasil prediksi terbaik didapatkan dengan 200 data yang telah dilakukan resample dan menggunakan SVM kernel RBF parameter C > 1, dan γ > 10. Percobaan tersebut menghasilkan nilai akurasi sebesar 92 %, precision 75 %, recall 71 %, dan f-measure 72.77 %..

Kata Kunci: ***Data Mining, Kanker Serviks, Penyakit, Prediksi, Support Vector Machine***

PREDICTION OF CERVICAL CANCER DIAGNOSIS BASED ON INFORMATION DEMOGRAPHIC, HABITS, AND MEDICAL RECORDS USING SUPPORT VECTOR MACHINE

Student Name	**: Siti Hawa Aminah**
NRP	**: 05211240000054**
Department	**: Information Systems FTIK-ITS**
Supervisors	**: Renny Pradina, S.T., M.T., SCJP**

ABSTRACT

Cervical cancer is one of the most chronic disease in the world. According to the Internationl Agency for Research on Cancer 2012, cervical cancer ranks third in disease that is often afflicted by women in the world, and ranks second in Indonesia. According to cancer experts, cervical cancer is one of the most preventable and most curable cancers of all cancer cases. One of the most common cervical cancer detection activities in Indonesia is the Pap Smear method. However, the availability of patient medical record history data is not accompanied by data extraction process to an information that can be useful for clinical decisions [1].

The concept of classification is part of the data mining technique that have the main work of doing predictive analysis. There have been many studies to make predictions and classifications in medical cases, as in the studies [2] *and* [3]. *Both studies produce a classification result with high accuracy rate, with an accuracy > 80 %. So it needs more research to get a model that is able to classify with minimal error rate and the model can be used to predict in the next data.*

This research use Cross Validation method to make the training process more accurate and get better prediction result. The

prediction experiments were done with SVM linear and RBF kernel. Use 27 attributes with 1 target attribute that is the result of patient's biopsy test. Total data that used on this research is 668 data and 200 resampled data.

The result of this research are the result of prediction performed on the dataset of patient 'Hospital Universitario de Caracas' in Caracas, Venezuela, using Support VectorMachine algorithm. The best prediction results were obtained with 200 resampled data and using SVM RBF kernel with parameters C > 1 and γ > 10. The experiments resulted in an accuracy of 92 %, precision 75 %, recall 71 % and f-measure 72.77%..

Keywords: ***Cervical Cancer, Data Mining, Disease, Prediction, Support Vector Machine***

KATA PENGANTAR

Puji syukur penulis haturkan kepada Allah SWT yang telah melimpahkan rahmat dan anugerah-Nya sehingga penulis dapat menyelesaikan Tugas Akhir dengan judul "Prediksi Diagnosa Kanker Serviks Berdasarkan Informasi Demografi, Kebiasaan, dan Rekam Medis Menggunakan Algoritma *Support Vector Machine"* sebagai salah satu syarat kelulusan pada Departemen Sistem Informasi, Fakultas Teknologi Informasi dan Komunikasi, Institut Teknologi Sepuluh Nopember. Semoga apa yang tertulis dalam buku Tugas Akhir ini dapat bermanfaat kepada para pembacanya, dan dapat memberikan kontribusi dalam perkembangan ilmu pengetahuan.

Dalam penyusunan Tugas Akhir ini tentunya sangat banyak bantuan yang penulis terima dari berbagai pihak baik dalam bentuk doa, semangat, arahan, kritik, saran, dan berbagai bantuan lainnya. tanpa mengurangi rasa hormat penulis secara khusus ingin menyampaikan ucapan terima kasih kepada:

1. Kedua orangtua penulis yang selalu memberikan dukungan materiil dan non-materiil demi terselesainya pengerjaan tugas akhir ini.
2. Ibu Renny Pradina K S.T, M.T, SCJP selaku dosen pembimbing yang telah sabar membimbing, memberi semangat, dan membantu penulis selama pengerjaan tugas akhir ini.
3. Bapak Faizal Johan Atletiko, S.Kom, M.T, dan Bapak Radityo Prasetianto Wibowo, S.Kom, M.Kom selaku dosen penguji yang telah memberikan masukan-masukan guna menyempurnakan Tugas Akhir ini.
4. Ibu Mahendrawati ER., S.T., M.Sc., Ph.D., dan Ibu Feby Artwodini, S.Kom., M.T., selaku dosen Departemen Sistem Informasi ITS yang selalu mengingatkan dan memberi semangat untuk penyelesaian Tugas Akhir ini.

5. Seluruh Dosen Departemen Sistem Informasi ITS yang telah memberikan ilmu pengetahuan yang bermanfaat dan pengalaman yang berharga bagi penulis.
6. Mas Fangky, Mas Fandy , dan Mbak Asmaul selaku saudara kandung dan saudara ipar penulis yang selalu mendukung untuk menyelesaikan Tugas Akhir
7. Dian dan Lintang, dua sahabat terbaik yang selalu mengingatkan, mendukung, dan membantu penulis untuk menyelesaikan Tugas Akhir.
8. Mbak Laily dan Imam, selaku mahasiswa yang berjuang bersama dalam mengerjakan Tugas Akhir.
9. Affandi, Prasetyo, Alden, dan Hendro selaku adik – adik 2013 dan 2014 yang siap memberikan masukan dan informasi untuk Tugas Akhir dan yudisium
10. 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

Halaman ini sengaja dikosongkan.

BAB I
PENDAHULUAN

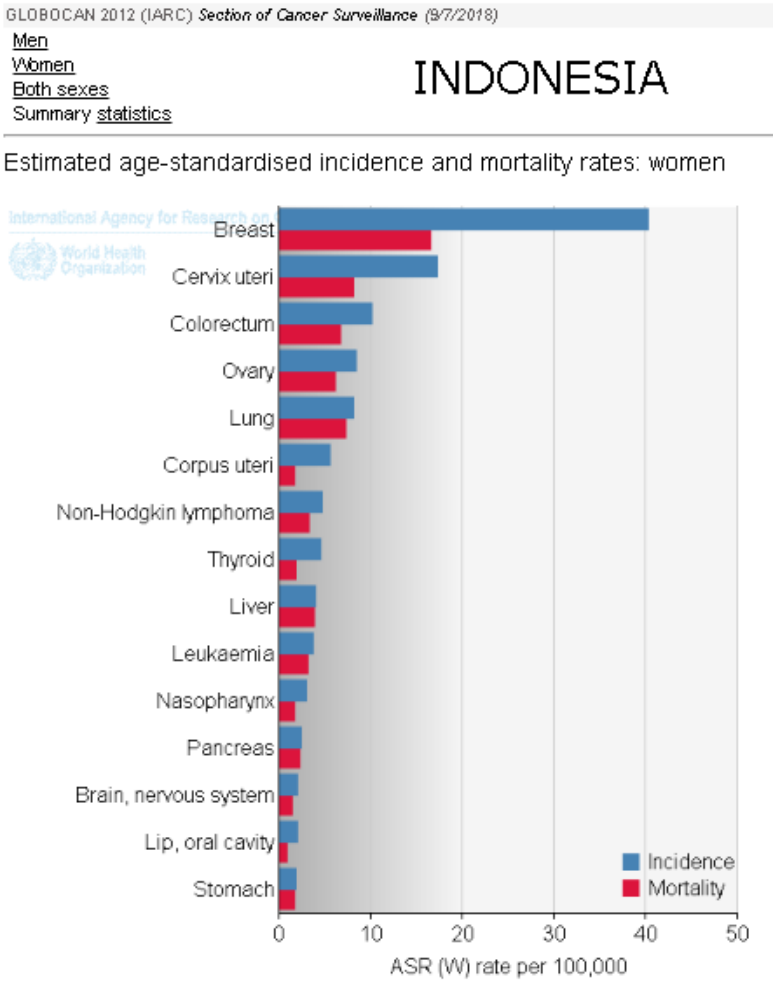
Pada bab pendahuluan akan diuraikan gambaran umum penelitian meliputi latar belakang masalah, perumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan relevansi terhadap pengerjaan tugas akhir.

1.1 Latar Belakang

Penyakit kanker masih menjadi penyebab utama kematian di seluruh dunia. Salah satu kanker yang dianggap sebagai penyakit paling mematikan yaitu kanker serviks. Kanker serviks merupakan kanker yang menyerang leher rahim, ditandai dengan pertumbuhan sel abnormal pada sel leher rahim. Kanker serviks banyak terdapat pada wanita Amerika Latin, Afrika, dan negara – negara berkembang lainnya di Indonesia, termasuk Indonesia. Menurut *International Agency for Research on Cancer* 2012, kanker serviks menempati peringkat ketiga pada penyakit yang sering diderita oleh wanita di seluruh dunia, dan menempati peringkat kedua di Indonesia.

Berdasarkan Data Riset Kesehatan Dasar 2013, prevalensi penyakit kanker pada penduduk semua umur di Indonesia sebesar 1,4% atau sekitar 347.792 orang. Dan penyakit kanker serviks merupakan penyakit kanker dengan prevalensi tertinggi di Indonesia pada tahun 2013, yaitu 0,8% atau sekitar 98.692 orang. [4]

Kanker serviks adalah kanker yang muncul pada leher rahim wanita. Jenis kanker ini dipicu oleh infeksi *Human Papilomavirus* (HPV). HPV merupakan sekumpulan virus yang menyebabkan kutil di bagian – bagian tubuh manusia. Beberapa faktor yang yang berisio tinggi terhadap terjadinya kanker serviks selain terinfeksi HPV adalah usia, usia saat pertama kali melahirkan, penggunaan alat kontrasepsi, pola hidup tidak sehat, sering berganti pasangan dalam berhubungan seksual, merokok, dan riwayat keluarga.



Gambar 1.1. Data penyakit yang diderita oleh wanita di Indonesia pada tahun 2012[1]

[1] *http://globocan.iarc.fr/Pages/fact_sheets_population.aspx*

Menurut para ahli kanker, kanker serviks adalah salah satu jenis kanker yang paling dapat dicegah dan paling dapat disembuhkan dari semua kasus kanker. Salah satu kegiatan deteksi dini kanker serviks yang paling umum di Indonesia adalah menggunakan metode *pap smear*. Namun tersedianya data histori rekam medis pasien tidak disertai dengan proses ekstraksi data menjadi sebuah informasi yang dapat berguna untuk keputusan klinis. [1]

Konsep klasifikasi merupakan bagian dari teknik data mining yang memiliki pekerjaan utama melakukan analis prediksi. Pada penelitian [1] teknik data mining digunakan untuk memprediksi penyakit kanker serviks menggunakan data rekam medis pasien ketika melakukan tes *pap smear*. Peneliti menggunakan hasil tes *pap smear* sebagai atribut target prediksi. Kemudian pada penelitian [5] algoritma yang digunakan untuk mengklasifikasi sel kanker dengan menerapkan *Genetic Algorithm* (GA) untuk seleksi fitur, kemudian mengklasifikasikan sel – sel sehat dan sel – sel kanker menggunakan algoritma *Support Vector Machine* (SVM). Fitur yang terbaik yang dipilih digunakan oleh SVM untuk klasifikasi pada dataset training untuk mengklasifikasikan sel. Dua penelitian tersebut menghasilkan klasifikasi dengan tingkat akurasi yang cukup tinggi, dengan nilai akurasi > 80 %. Sehingga diperlukan lebih banyak penelitian untuk mendapatkan suatu model yang mampu mengklasifikasikan dengan tingkat kesalahan minimal dan model yang dihasilkan dapat digunakan untuk melakukan prediksi data berikutnya.

Selain dua penelitian diatas, penelitan dalam tugas akhir ini menggunakan dataset dari paper yang berjudul '*Trasnfer Learning with Partial Observability Applied to Cervical Cancer Screening*'. Penelitian tersebut melakukan prediksi kanker serviks berdasarkan faktor risiko dari dataset responden pasien rumah sakit serta dataset yang telah diverifikasi oleh para ahli, dengan tujuan mencari metode *transfer learning* yang lebih baik dari metode yang lain. Penelitian tersebut membandingkan hasil dari *gain* yang diperoleh saat pengukuran *signed Area*

Under the gain Curve (sAUC), dan mendapatkan hasil yang positif. Metode yang dipakai pada paper tersebut mendapatkan hasil yang lebih baik dari dua metode yang ada.

Penulis menggunakan dataset yang digunakan oleh penelitian [6]. Penelitian tersebut menggunakan seluruh dataset (858 baris data), baris data yang terdapat *missing value* diisi dengan hasil rata – rata dari atribut / variabel. Pada penelitian [6] fokus kepada metode *transfer learning* yang akan ditawarkan, sehingga tidak membahas mengenai dataset yang tidak seimbang, serta tidak ditampilkan bagaimana keakuratan, ketepatan, dan keberhasilan percobaan prediksi pada data. Pada tugas akhir ini, penulis akan mencoba bagaimana agar prediksi bisa menghasilkan performa yang baik di semua sisi, tidak hanya bagus dari sisi akurasi saja.

Untuk metode prediksi yang digunakan, penelitian tugas akhir ini menggunakan rujukan dari penelitian [2], yang melakukan prediksi kanker serviks dengan metode MLP, *Bayes Net*, dan k-NN. Namun penulis akan menggunakan metode *machine learning* yang lain yaitu SVM, sehingga hasil akhir dari penelitian diharapkan agar dapat membandingkan dengan hasil prediksi yang telah dilakukan sebelumnya.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, maka perumusan permasalahan yang akan diselesaikan dalam tugas akhir ini adalah sebagai berikut:

1. Bagaimana melakukan prediksi diagnosa kanker serviks dengan menggunakan algoritma *Support Vector Machine* (SVM) dengan kernel *linear* dan RBF?
2. Bagaimana hasil pengukuran performa dari prediksi yang dihasilkan menggunakan algoritma *Support Vector Machine* (SVM)?

1.3 Batasan Masalah

Berdasarkan permasalahan yang dihadapi, terdapat batasan masalah yang ditetapkan sebagai fokus pengerjaan tugas akhir. Batasan yang ditetapkan dalam tugas akhir ini adalah sebagai berikut:

1. Data yang digunakan adalah dataset pasien *'Hospital Universitario de Caracas'* di Caracas, Venezuela yang didapatkan dari UCI *Machine Learning Repository.*
2. Atribut yang digunakan adalah
    a. *Age*
    b. *Number of sexual partners*
    c. *First sexual intercourse*
    d. *Number of pregnancies*
    e. *Smokes*
    f. *Smokes (years)*
    g. *Smokes (packs/years)*
    h. *Hormonal Contraceptives*
    i. *Hormonal Contraceptives (years)*
    j. *IUD*
    k. *IUD (years)*
    l. *STDs*
    m. *STDs (number)*
    n. *STDs : condylomatosis*
    o. *STDs : cervical condylomatosis*
    p. *STDs : vaginal condylomatosis*
    q. *STDs : vulvo-perineal condylomatosis*
    r. *STDs : syphilis*
    s. *STDs : pelvic inflammatory disease*
    t. *STDs : genital herpes*
    u. *STDs : molluscum contagiosum*
    v. *STDs : AIDS*
    w. *STDs : HIV*
    x. *STDs : Hepatitis B*
    y. *STDs : HPV*
    z. *STDs : Number of diagnosis*
    aa. *Biopsy*

1.4 Tujuan

Berdasarkan perumusan masalah yang disebutkan sebelumnya, tujuan yang ingin dicapai melalui tugas akhir ini adalah :

1. Menerapkan teknik data mining yaitu prediksi data menggunakan algoritma SVM dengan kernel *linear* dan RBF.
2. Membandingkan pengukuran performa hasil prediksi.

1.5 Manfaat

Manfaat yang dapat diperoleh dari pengerjaan tugas akhir ini adalah mengimplementasikan teknik data mining dengan algoritma SVM dalam memprediksi diagnosa kanker serviks.

1.6 Relevansi

Tugas Akhir ini layak disebut sebagai salah satu bentuk penelitian studi pada keilmuan Sistem Informasi dikarenakan keterkaitan fokus penelitian ini dengan area penelitian dalam bidang Sistem Informasi dalam lingkup penelitian Laboratorium Akuisisi Data dan Diseminasi Informasi, karena penelitian ini menerapkan beberapa mata kuliah yang berhubungan dengan laboratorium ini. Adapun mata kuliah terkait, di antaranya:

- Penggalian Data dan Analitik Bisnis
- Sistem Pendukung Keputusan
- Sistem Cerdas

BAB II
TINJAUAN PUSTAKA

Pada bab tinjuan pustaka akan menjelaskan mengenai penelitian sebelumnya yang terkait dengan tugas akhir dan membahas dasar teori yang perlu dipahami untuk dijadikan acuan atau landasan dalam pengerjaan tugas akhir ini.

2.1 Penelitian Sebelumnya

Rujukan penelitian pertama pada penelitian ini adalah paper yang berjudul "*Transfer Learning with Partial Observability Applied to Cervical Cancer Screening"* oleh Kelwin Fernandes, dkk. Penelitian tersebut melakukan prediksi kanker serviks berdasarkan faktor risiko dari dataset 858 responden pasien *'Hospital Universitario de Caracas'* di Caracas, Venezuela serta dataset yang telah diverifikasi oleh para ahli kanker serviks, dengan tujuan mencari metode *transfer learning* yang lebih baik dari metode yang lain. Data yang digunakan pada penelitian ini terdapat beberapa *missing value*, namun diisi dengan nilai rata – rata dari atribut / variabel tersebut. Penelitian tersebut membandingkan hasil dari *gain* yang diperoleh saat pengukuran *signed Area Under the gain Curve* (sAUC), dan mendapatkan hasil yang positif [6]. Pada tugas akhir ini akan menggunakan salah satu dataset dari penelitian yang dilakukan oleh Kelwin Fernandes, yaitu dataset dari 858 pasien *'Hospital Universitario de Caracas'*.

Rujukan penelitian kedua adalah paper yang ditulis oleh G. Ravi Kumar, dkk yang berjudul "*An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets"*. Penelitian fokus pada mengkombinasikan seleksi fitur berdasarkan *Genetic Algorithm* dan SVM dari klasifikasi penyakit. Penelitian menggunakan dataset dari berbagai jenis penyakit, kemudian melakukan evaluasi metrik (*Precision, Recall, dan F-score*) dari masing – masing dataset. Akurasi terbaik dihasilkan dengan menggunakan metode GA-SVM daripada SVM [3]. Pada

penelitian tugas akhir ini akan menggunakan salah satu metode, yaitu SVM, serta melakukan perhitungan *precision, recall,* dan *f-measure*.

Rujukan penelitian ketiga adalah paper berjudul "Klasifikasi Hasil Pap Smear Test sebagai Upaya Pencegahan Sekunder Penyakit Kanker Servis di Rumah Sakit "X" Surabaya Menggunakan *Piecewise Polynomial Smooth Support Vector Machine* (PPSSVM)" oleh Mukti Ratna Dewi dan Santi Wulan Purnami. Penelitian ini menggunakan data sekunder yang diperoleh dari Rumah Sakit "X" Surabaya bagian Riset dan Pengembangan tahun 2010 untuk membentuk model klasifikasi dengan metode SSVM dan PPSSVM. Dengan kesimpulan metode PPSSVM memiliki performa lebih baik dari SSVM. Namun metode PPSSVM tidak efisien dalam waktu komputasi [7].

Rujukan penelitian keempat adalah paper yang ditulis oleh Muhammad Fahri Unlersen dan Kadir Sabanci yang berjudul "*Determining Cervical Cancer Possibility by Using Machine Learning Methods".* Penelitian ini melakukan klasifikasi terhadap dataset dari 858 pasien *'Hospital Universitario de Caracas'* menggunakan *Multilayer Perceptron* (MLP), *Bayes Net* dan k-NN. Pada penelitian ini menggunakan 66% data untuk data *training*, yaitu 566 baris data, dan 292 sisanya untuk data *testing.* Akurasi tertinggi diperoleh yaitu 97.26% menggunakan k-NN. [2] Penelitian yang akan dilakukan pada tugas akhir tidak jauh berbeda dari penelitian ini, namun penulis akan menggunakan algoritma SVM untuk membandingkan dari hasil penelitian [2]

Untuk mengetahui lebih jelas dan detail terhadap penelitian sebelumnya, dapat dilihat pada Tabel 2.1 :

Tabel 2.1. Perbandingan Penelitian Sebelumnya

No	Judul	Penulis (Tahun)	Keterangan
1	*Transfer Learning with Partial Observability Applied to Cervical Cancer Screening*	Kelwin Fernandes, Jaime S. Cardoso, Jessica Fernandes. (2017)	• Bertujuan untuk mencari metode *transfer learning* yang lebih baik dari metode sebelumnya. • Menggunakan 858 data pasien rumah sakit dan data yang telah diverifikasi oleh ahli kanker serviks • Penilaian dilakukan dengan membandingkan hasil *signed Area Under the gain Curve* (sAUC)

No	Judul	Penulis (Tahun)	Keterangan
2	*An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets*	G. Ravi Kumar, Dr. G. A. Ramachandra, dan K. Nagamani. (2014)	• Menggunakan dataset beberapa penyakit • Melakukan klasifikasi berdasarkan GA dan SVM • Melakukan evaluasi metrik (*precision, recall, f-score)*
3	Klasifikasi Hasil Pap Smear Test sebagai Upaya Pencegahan Sekunder Penyakit Kanker Servis di Rumah Sakit "X" Surabaya Menggunakan *Piecewise Polynomial Smooth Support Vector Machine* (PPSSVM)"	Mukti Ratna Dewi dan Santi Wulan Purnami (2015)	• Menggunakan data sekunder dari rumah sakit "X" Surabaya bagian Riset dan Pengembangan tahun 2010 • Melakukan klasifikasi dengan metode SSVM dan PPSSVM

No	Judul	Penulis (Tahun)	Keterangan
4	*Determining Cervical Cancer Possibility by Using Machine Learning Methods*	Muhammad Fahri Unlersen dan Kadir Sabanci (2017)	• Menggunakan 858 data pasien rumah sakit seperti penelitian [6] • Melakukan klasifikasi dengan MLP, *Bayes Net*, dan k-NN. • 66% data digunakan sebagai data training

2.2 Dasar Teori

Pada subbab ini akan dijelaskan mengenai teori – teori yang digunakan untuk referensi dalam pengerjaan tugas akhir.

2.2.1 Kanker Serviks

Kanker servis adalah kanker yang terjadi pada serviks uterus, suatu daerah pada organ reproduksi wanita yang merupakan pintu masuk ke arah rahim yang terletak antara rahim (*uterus*) dengan liang senggama (*vagina*). Kanker serviks terbentuk sangat perlahan dan sangat sulit terdeteksi di awal kecuali terjadi infeksi pada fisik. Pertama, tumbuhnya siklus sel kanker yang ada dalam tubuh berubah dari keadaan normal menjadi sekumpulan sel pra kanker, kemudian berkembang menjadi sel kanker. Perubahan selkanker ini terjadi secara bertahap, dan memerlukan waktu bertahun – tahun. Namun

tidak jarang pertumbuhan sel kanker ini berlangsung secara cepat. Hal tersebut disebabkan oleh daya tahan tubuh setiap orang berbeda. [8]

Penelitian World Health Organization (WHO) menyebutkan bahwa di seluruh dunia terdapat 490.000 kasus kanker serviks dan mengakibatkan 240.000 kematian tiap tahunnya. 80% terjadi di Asia.

Menurut Departemen Kesehatan RI, saat ini jumlah penderita baru kanker serviks berkisar 90 – 100 kasus per 100.000 penduduk dan setiap tahun terjadi 40.000 kasus kanker serviks. Kejadian kanker serviks akan mempengaruhi hidup dari penderita dan keluarga serta mempengarhui sektor pembiayaan kesehatan oleh pemerintah. Sehingga peningkatan upaya penanganan kanker serviks sangat diperlukan oleh pihak yang terlibat terutama dalam pencegahan dan deteksi dini kanker.

Pada umumnya, penderita kanker serviks belum terdeteksi gejala pada fisik. Jika telah terjadi kanker invasif, gejala yang paling umum adalah pendarahan saat berhubungan intim, dan keputihan. Pada stadium lanjut, gejala dapat berkembang menjadi nyeri pinggang atau perut bagian bawah. Deteksi dini kanker serviks dapat dilakukan dengan pemeriksaan klinik. Pemeriksaan klinik meliputi inspeksi, kolposkopi, biopsi serviks, sistoskopi, rektoskopi, USG, BNO – IVP, foto toraks, dan CT scan. [8]

Penelitian ini menggunakan target atribut hasil tes biopsi. Biopsi adalah pengambilan bagian kulit, jaringan, atau organ tubuh untuk diperiksa di laboratorium. Tujuan dari biopsi adalah untuk mengetahui apakah suatu jaringan mengandung sel – sel kanker atau sel – sel abnormal lainnya. Biopsi serviks adalah tindakan ginekologi untuk mengambil sampel jaringan dari serviks atau leher rahim.

2.2.2 Faktor Risiko Kanker Serviks

Penyebab kanker serviks adalah *Human Papiloma Virus* (virus HPV) sub tipe onkogenik, terutama sub tipe 16 dan 18.

Kanker serviks menyerang wanita yang pernah atau sedang aktif secara seksual. Pada umumnya kanker serviks menyerang wanita berusia 33 – 55 tahun. Namun tidak menutup kemungkinan wanita yang lebih muda dapat menderita penyakit ini jika terdapat faktor risikonya.

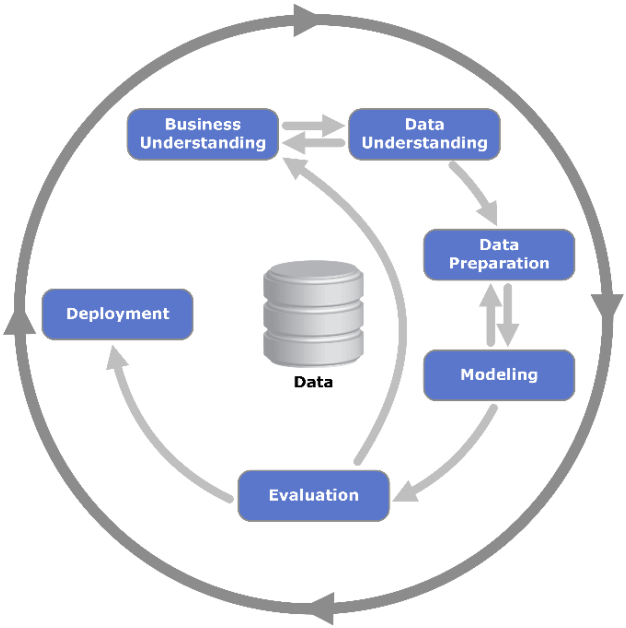
Beberapa faktor risiko yang kanker serviks :

- usia
- usia pertama kali melakukan hubungan seksual
- pasangan seksual yang berganti – ganti
- kurang menjaga kebersihan genital
- merokok
- riwayat penyakit kelamin
- penggunaan kontrasepsi
- gangguan imunitas

Selain faktor yang telah disebutkan, terlambat menikah juga merupakan faktor risiko kanker serviks karena golongan wanita ini akan terus – menerus mengalami ovulasi tanpa jeda, sehingga rangsangan terhadap endometrium terjadi terus – menerus dan dapat membuat sel – sel endometrium berubah sifat menjadi kanker. [9]

2.2.3 Data Mining

Data Mining merupakan teknologi baru yang berguna untuk membantu perusahaan – perusahaan dalam menemukan informasi penting dari gudang data yang dimiliki. Beberapa aplikasi data mining fokus pada prediksi, yaitu meramalkan apa yang akan terjadi dalam situasi baru dari data yang menggambarkan apa yang terjadi di masa lalu. Data mining merupakan bagian dari *Knowledge Discovery Data* (KDD) yang merupakan proses ekstraksi informasi yang berguna, tidak diketahui sebelumnya, dan tersembunyi dari data, serta mengembangkan model yang digunakan untuk memahami fenomena dari analisis data dan prediksi.



Gambar 2.1. Tahapan CRISP-DM[2]

Standar proses yang digunakan dalam data mining adalah kerangka kerja CRISP-DM. Dalam CRISP-DM, data mining dilihat sebagai proses keseluruhan dari komunikasi masalah bisnis hingga penerapan model. [10], proses CRISP-DM terdiri dari enam tahap, yaitu :

1. Tahap *Business Understanding*
   Tahap *business understanding* disebut juga sebagai tahap *research understanding*. Dalam tahap ini ditentukan tujuan dan *requirement* secara detail pada keseluruhan penelitian, merumuskan masalah data mining, dan menyiapkan strategi awal untuk mencapai tujuan yang telah ditentukan.

[2] http://eprints.dinus.ac.id/14738/1/[Materi]_Cross-Industry_Standard_Process_for_Data_Mining.pdf

2. Tahap *Data Understanding*
   Pada tahap ini mulai dilakukan proses pengumpulan data, analisis data, evaluasi kualitas data, dan memilih subset yang mungkin mengandung pola yang akan ditindaklanjuti.
3. Tahap *Data Preparation*
   Pada tahap ini, data akhir yang akan digunakan pada tahap berikutnya mulai disiapkan, memilih kasus dan atribut yang sesuai dengan analisis yang akan dilakukan, melakukan transformasi pada atribut tertentu jika diperlukan, dan membersihkan data mentah sehingga siap untuk digunakan sebagai alat pemodelan.
4. Tahap *Modelling*
   Tahap ini dilakukan untuk memilih dan menerapkan teknik pemodelan, menentukan *tools* data mining yang digunakan, serta menentukan parameter dengan nilai yang optimal.Pada tahap ini juga disebut sebagai tahap *learning* karena data training dilatih oleh model yang dipilih.
5. Tahap *Evaluation*
   Pada tahap evaluation, menilai model agar dapat menentukan apakah model tersebut dapat memenuhi tujuan bisnis.
6. Tahap *Depoloyment*
   Tahap ini merupakan tahapan untuk rencana penggunaan model yang telah disetujui.

2.2.4 *Imbalanced Data*

Imbalanced data atau data tidak seimbang mengacu pada masalah klasifikasi data dimana data pada kelas yang ada tidak seimbang. Sebagai contoh, kita mempunyai sebuah dataset yang memiliki 1000 baris dan dua kelas A dan B. Kelas A memiliki 900 baris sedangkan label kelas B hanya 100 baris. Dengan jumlah data yang tidak seimbang tersebut, ketika melakukan klasifikasi / prediksi, kemungkinan akan mendapatkan nilai

akurasi sebesar 90%. Karena 90% dari data terdapat pada satu label kelas [11]. Hasil akurasi bukanlah cara yang baik untuk mengukur keberhasilan klasifikasi / prediksi yang diterapkan pada *imbalanced data*. Kelas mayoritas akan memiliki nilai akurasi yang lebih tinggi daripada kelas minoritas.

Beberapa cara untuk mengatasi *imbalanced data* :

- Menentukan parameter yang berbeda

  Pada saat melakukan klasifikasi dengan SVM, melakukan uji coba dengan parameter C dan γ yang beragam.

- Menggunakan lebih dari satu pengukuran performa

  Menggunakan evaluasi metrik seperti *precision, recall, f-measure*.

- *Resampling* data

  Resampling data dapat dilakukan dengan *undersampling / downsampling* (mengurangi) dan *oversampling / upsampling* (menambah dengan menduplikat kelas yang minoritas secara random)

- Mengunakan *Cross-Validation*

  Teknik validasi silang yang akan dibahas pada sub judul 2.2.6.

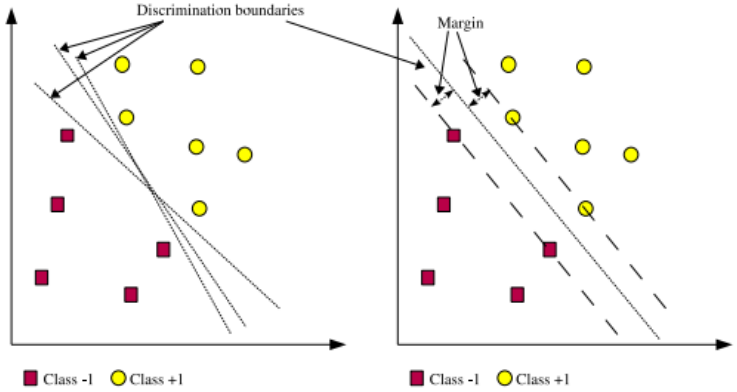
2.2.5 Support Vector Machine

Support Vector Machine (SVM) adalah salah satu metode *supervised learning* yang digunakan untuk kebutuhan klasifikasi, regresi, dan pendeteksian *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 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 *supervised learning*, akan diperoleh fungsi yang menggambarkan bentuk ketergantungan input dan *output*nya

2.2.5.1 Kernel Linear

Kernel linear digunakan ketika data yang akan diklasifikasikan dapat terpisah dengan sebuah garis / *hyperplane*. *Hyperplane* merupakan batas yang memisahkan data antara kelas 1 dengan kelas lainnya. Pada kernel linear, parameter yang digunakan adalah parameter C (*cost*).



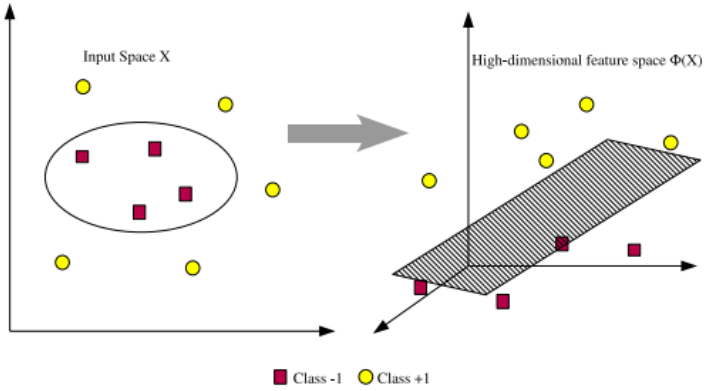
Gambar 2.2. Visualisasi Kernel Linear[3]

2.2.5.2 Kernel RBF

Kernel *Radial Basis Function* (RBF) termasuk dalam kernel non – linear, yaitu ketika data yang digunakan hanya dapat dipisahkan dengan garis lengkung atau sebuah bidang pada ruang dimensi tinggi. Pada kernel RBF menggunakan dua parameter, yaitu C (cost) dan γ (gamma). Secara intuitif,

[3] https://asnugroho.wordpress.com/2007/02/05/pengantar-support-vector-machine/

parameter C akan melihat kesalahan dalam klasifikasi dari contoh data *training*. Parameter γ mendefinisikan seberapa jauh pengaruh dari contoh *training* tunggal tercapai, dengannilai rendah berarti ‘jauh’ dan nilai tinggi berarti ‘dekat’.

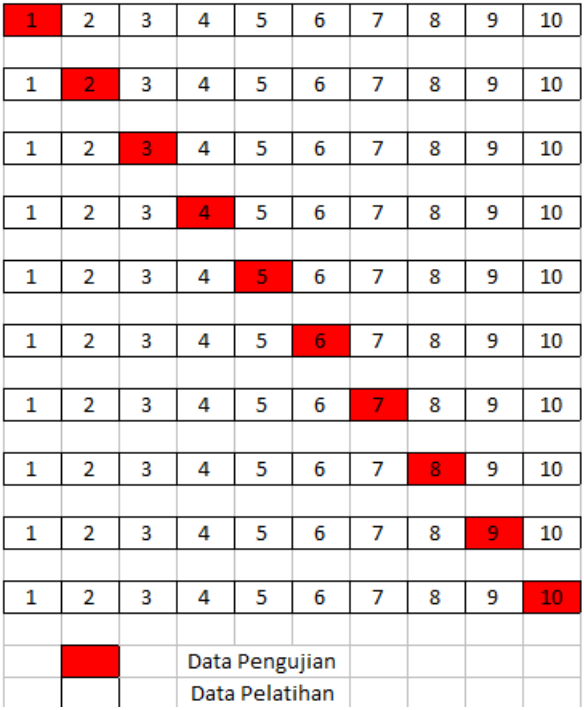


Gambar 2.3. Visualisasi Kernel RBF[4]

2.2.6 Cross Validation

Metode validasi silang (*cross validation*) merupakan metode yang cukup populer untuk evaluasi. Pada metode ini, data dibagi menjadi dua bagian, yaitu data training dan data testing. Kemudian data yang diuji akan dilakukan proses silang, dimana data training menjadi data testing, atau sebaliknya. Metode *cross validation* akan menghindari tumpang tindih pada data testing, dan mengurangi waktu komputasi dengan tetap menjada akurasi. Metode ini juga biasa disebut *k-fold.* Nilai k yang paling sering digunakan dan direkomendasikan adalah 10 untuk memilih model terbaik. Dalam *10-fold cross validation*, data akan dibagi menjadi 10 lipatan / partisi. Selanjutnya dilakukan eksperimen menggunakan data yang telah dipartisi, dengan posisi partisi data testing yang berbeda.

[4] https://asnugroho.wordpress.com/2007/02/05/pengantar-support-vector-machine/



Gambar 2.4. ***10-fold Cross Validation***

2.2.7 Pengukuran Performa

Tingkat akurasi deteksi dari suatu model merupakan hal penting untuk diperhatikan. Penilaian yang paling umum digunakan dalam data mining adalah Akurasi, *Precision*, *Recall*, dan *F-Measure*. Keempat ukuran performa tersebut memiliki nilai pada range 0 hingga 1. Semakin mendekati nilai 1, maka performa akan semakin baik.

Confusion matrix merupakan alat analisis yang berguna untuk menangkap hasil evaluasi dan sebagai basis dalam perhitungan performa lain. Co*nfusion matrix* menghitung frekuensi pada tiap kemungkinan yang keluar pada prediksi model. [12]

Untuk prediksi dengan fitur biner (2 level), akan terdapat 4 luaran hasil sebagai berikut :

- *True Positive* (TP)
  Nilai pada data testing positif, dan prediksi positif
- *True Negative* (TN)
  Nilai pada data testing negatif, dan prediksi negatif
- *False Positive* (FP)
  Nilai pada data testing negatif, dan prediksi positif
- *False Negative* (FN)
  Nilai pada data testing positif, dan prediksi negatif

Tabel 2.2. Tabel Confusion Matrix

		Prediksi	
		Positif	**Negatif**
Data Testing (Aktual)	**Positif**	*True Positive*	*False Negative*
	Negatif	*False Positive*	*True negative*

a. *Accuracy*
   Merupakan presentase dari total data yang terprediksi dengan benar.

$$Accuracy = \frac{True\ Positive + True\ Negative}{Total\ Data}$$

b. *Precision*
   Precision merupakan nilai ketepatan prediksi dengan benar. Perhitungan *precision* dilakukan dengan membandingkan jumlah prediksi benar dengan jumlah seluruh prediksi

$$\text{Precision} = \frac{True\ Positive}{False\ Positive + True\ Positive}$$

c. *Recall*
Recall merupakan nilai perbandingan ketepatan prediksi benar dengan jumlah seluruh prediksi yang seharusnya benar.

$$Recall = \frac{True\ Positive}{False\ Negative + True\ Positive}$$

d. *F-Measure*
F-Measure adalah perhitungan yang digunakan untuk menggabungkan nilai *Precision* dan *Recall. F-Measure* menunjukkan nilai keseimbangan dari *Precision* dan *Recall*.

$$F\text{-}Measure = \frac{2\ x\ (Precission\ x\ Recall)}{Precission + Recall}$$

Untuk melihat bagaimana performa dari prediksi yang telah dilakukan, sebaiknya tidak dilihat dari satu nilai saja, tetapi dari semua nilai akurasi, *precision, recall,* dan *f-measure*. Jika hasil pengukuran akurasi, *precision, recall,* dan *f-measure* sama – sama memiliki nilai yang tinggi, dapat dipastikan bahwa model prediksi yang dihasilkan sangat baik. Beberapa perbedaan antara akurasi dan *precision* dapat dilihat pada Tabel 2.3.

Semua pengukuran selalu menghadapi ketidakpastian dalam pengukuran. Ketidakpastian adalah keraguan yang muncul pada hasil setiappengukuran. *Error* adalah perbedaan antara nilai hasil pengukuran dengan nilai sebenarnya. Sedangkan ketidakpastian adalah kuantifikasi dari keraguan tentang hasil pengukuran.

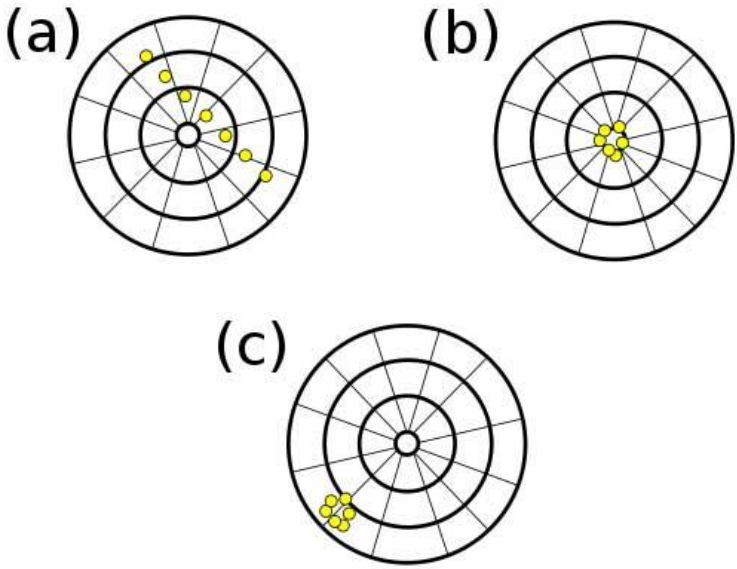
Tabel 2.3. Perbeedaan Akurasi dan *Precision*

	Akurasi	***Precision***
Pengertian	Mengacu pada level kesepakatan antara pengukuran aktual dan pengukuran absolut	Mengartikan level keberagaman yang terletak pada nilai beberapa pengukuran dari faktor yang sama
Menggambarkan	Seberapa dekat hasil dengan nilai standar	Seberapa dekat hasil dengan yang lain
Derajat	Derajat kecocokan	Derajat reprodusibilitas
Faktor	Faktor tunggal	Banyak faktor
Pengukuran dari	Perkiraan statikal	Keberagaman statikal
Terkait dengan	Kesalahan sistematis	Kesalahan acak

Kesalahan sistematis cenderung menggeser semua pengukuran secara sistematis, sehingga nilai rata – rata secara konstan bergeser atau bervariasi dalam cara yang dapat diprediksi.

Kesalahan acak adalah komponen dari kesalahan total yang dalam perjalanan dari sejumlah pengukuran, bervariasi dalam cara yang tak terduga. Kesalahan acak dapat terjadi karena berbagai alasan :

- Kurangnya kepekaan (sensivitas) *tools*.
  Tools yang tidak mampu merespon atau menunjukkan perubahan dalam beberapa kuantitas yang terlalu kecil
- Kebisingan (*noise)* dalam pengukuran
  Gangguan asing yang tak terduga dan tidak bisa sepenuhnya dihitung
- Definisi tidak tepat
  Sulit untuk menentukan dimensi sebuah obyek.



Gambar 2.5. Perbandingan Akurasi VS *Precision*[5]

Gambar 1.1. Data penyakit yang diderita oleh wanita di Indonesia pada tahun 2012 Gambar 2.5 merupakan ilustrasi dari perbandingan antara akurasi dan *precision.*

(a) Akurasi rendah, *precision* rendah. Tidak akurat dan tidak tepat.
(b) Akurasi tinggi, *precision* tinggi. Akurat dan tepat.
(c) Akurasi rendah, *precision* tinggi. Tepat, namun tidak akurat.

Untuk menghasilkan model prediksi yang baik, nilai masing – masing pengukuran akurasi, *precision, recall* dan *f-*

[5] https://www.tentorku.com/ketidakpastian-kesalahan-akurasi-dan-presisi/

measure harus tinggi, dan seimbang (tidak ada yang berbeda jauh lebih rendah atau lebih tinggi dari yang lain)

2.2.8 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.

Bahasa Python menjalankan perintah secara berurutan dan memiliki sistem indentasi yaitu memisahkan blok program dengan susunan indentasi. Pengguna Python cukup menggunakan spasi sebagai pemisah blok program. [13]

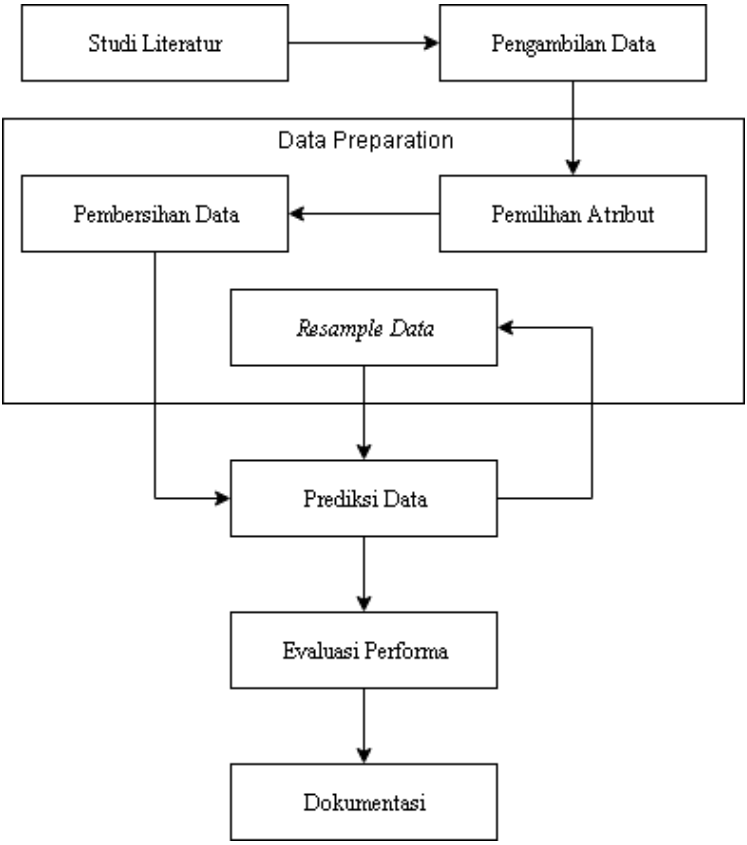
Pemrograman bahasa python adalah pemrogram gratis atau *freeware*, sehingga dapat dikembangkan, dan tidak ada batasan dalam penyalinannya dan distribusi. Terdapat beberapa pelayanan yang disediakan lengkap dengan *source code, debugger* dan *profiler, interfac*e, fungsi sistem, GUI, dan basis datanya.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

2.2.9 Anaconda

Anaconda merupakan open data science platform yang didukung oleh Python. Platform ini merupakan distribusi Python yang berisikan banyak software packages seperti co*nda package* dan *environment manager* untuk analisis data, ilmu data, dan komputasi ilmiah. [14] Anaconda memiliki versi *open source* dengan peforma terbaik dan memiliki lebih dari 720 packages untuk Python. Dengan menggunakan Anaconda, pengguna tidak perlu melakukan instalasi *packages*.

BAB III
METODOLOGI

Pada bab metodologi akan dijelaskan mengenai gambaran metode dan alur pengerjaan tugas akhir. Bab ini menjadi acuan dalam pengerjaan tugas akhir hingga proses pengerjaan menjadi sistematis.



Gambar 3.1. Tahapan Pelaksanaan Tugas Akhir

3.1 Studi Literatur

Tugas akhir dilakukan dengan menggali informasi dari penelitian yang pernah dilakukan sebelumnya. Studi literatur yang dilakukan adalah dengan mencari berbagai referensi seperti buku pustaka, penelitian sebelumnya, dan dokumen yang berkaitan dengan penelitian tugas akhir. Untuk mendukung latar belakang dan perumusan masalah sesuai topik yang dipilih, penulis melakukan tinjauan pustaka dimulai dari studi penelitian sebelumnya yang berhubungan dengan suatu prediksi yang menggunakan algoritma SVM.

Beberapa teori yang dipelajari yang berkaitan dengan penelitian tugas akhir adalah :

- Pengertian Kanker serviks
- Faktor risiko kanker serviks
- Konsep ata mining
- Pengertian i*mbalanced data*
- Teknik klasifikasi dan prediksi menggunakan *Support Vector Machine* (Kernel *linear* dan RBF)
- Melakukan training dengan metode c*ross validation*
- Pengukuran performa yang akan digunakan dalam mengevaluasi model prediksi
- Bahasa pemrograman python dan *library* yang digunakan
- Implementasi kode pada Anaconda Jupyter

3.2 Pengambilan Data

Tahapan selanjutnya adalah pengambilan data yang akan digunakan pada penelitian tugas akhir. Penulis mengambil data dari UCI *Machine Learning Repository*, yaitu dataset 858 pasien di *Hospital Universitario de Caracas*, Caracas, Venezuela.

3.3 Pemilihan Atribut

Setelah mendapatkan data, melakukan pemilihan atribut. Pemilihan atribut yang dilakukan adalah memilih atribut yang lengkap (isian dari atribut tidak ada dan banyak *missing value*). Untuk target atribut diagnosa kanker serviks pada data asli terdapat empat target yaitu hasil diagnosa menurut tes *Hinselmann*, tes *Schiller*, tes *Citology*, dan tes *Biopsy*. Penulis menggunakan hasil tes *Biopsy* untuk target atribut pada penelitian ini.

3.4 Pembersihan Data

Tahapan pembersihan data yaitu menghilangkan beberapa baris data yang tidak lengkap pada isian atribut. Dari sumber pengambilan data disebutkan bahwa terdapat beberapa pasien rumah sakit yang tidak mengisi beberapa atribut karena terkait dengan privasi pasien tersebut.

3.5 Prediksi Data

Pada tahapan ini data training akan dilatih oleh model yang dipilih. Training data juga dilakukan dengan menggunakan *K-Fold Cross Validation* agar proses training lebih akurat dan menghasilkan prediksi yang lebih baik. Kemudian akan dilakukan testing dari hasil model training untuk data target prediksi. Proses klasifikasi menggunakan SVM kernel *linear* dan RBF. Percobaan prediksi yang dilakukan adalah sebagai berikut :

- Prediksi 668 data dengan SVM kernel *linear*
- Prediksi 668 data dengan SVM kernel RBF
- Prediksi 200 data *resampled* dengan SVM kernel *linear*
- Prediksi 200 data *resampled* dengan SVM kernel RBF

3.6 Resample Data

Tahapan *resample data* dilakukan karena data yang didapatkan tidak seimbang antara label kelas satu dengan yang lain. Pada tahap ini akan dilakukan *downsampling* pada data label kelas mayoritas yaitu kelas negatif kanker, dari 623 menjadi 100. Dan melakukan *upsampling* pada data label kelas minoritas yaitu kelas positif kanker, dari 45 menjadi 100.

3.7 Evaluasi Performa

Merupakan interpretasi hasil prediksi yang telah dilakukan. Evaluasi dilakukan dengan membandingkan nilai akurasi, *precision*, *recall*, dan *f-measure* dari dua kernel *linear* dan RBF.

3.8 Dokumentasi

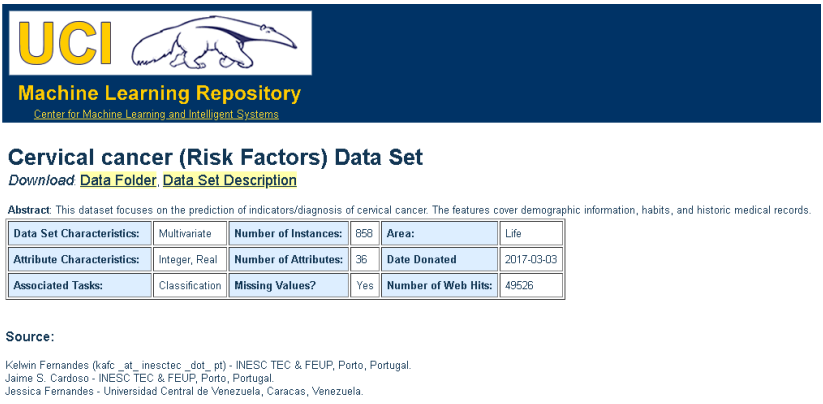
Dokumentasi yang dilakukan adalah penulisan buku Tugas Akhir. Penulisan buku ini dilakukan bersamaan dengan tahapan penelitian sebelumnya.

BAB IV
PERANCANGAN

Pada bab ini membahas terkait alur perancangan terkait beberapa hal yang diperlukan dalam proses pembuatan aplikasi sesuai dengan alur yang dijelaskan pada bab metodologi. Adapun perancangan ini diperlukan sebagai panduan dalam melakukan penelitian tugas akhir, yang dijelaskan sebagai berikut :

4.1 Pengambilan Data

Data merupakan salah satu komponen penting untuk melakukan penelitian prediksi diagnosis kanker serviks menggunakan algoritma *Support Vector Machine*. Data dalam penelitian ini merupakan dataset menggunakan format file ekstensi *Comma Separated Value* (.csv) dari pasien '*Hospital Universitario de Caracas'* di Caracas, Venezuela. Data tersebut diunduh pada website UCI *Machine Learning Repository*.



Gambar 4.1. UCI *Machine Learning Repository*

4.2 Pemilihan Atribut

Pemilihan atribut dilakukan untuk mempertimbangkan dan menyeleksi atribut data yang akan digunakan pada penelitian. Penelitian ini menggunakan atribut target hasil tes *biopsy* pasien dengan dua label yaitu 1 (positif kanker) dan 0 (negatif kanker). Tiga atribut target lain yaitu tes menurut *Hinselmann*, tes *Schiller*, dan tes *Cytology* tidak digunakan. Beberapa atribut yang memiliki banyak *missing value* tidak digunakan, sehingga penulis menggunakan 27 dari 36 atribut yang ada. 27 atribut inilah yang akan digunakan untuk tahapan data *preparation* selanjutnya.

27 atribut yang digunakan adalah :

a. *Age*
Menunjukkan usia pasien. Dari data yang didapatkan, usia pasien memiliki rentang 13 – 84 tahun.

b. *Number of sexual partners*
Merupakan jumlah pasangan yang pernah berhubungan seksual dengan pasien. Atribut ini memiliki rentang nilai 1 – 28 kali.

c. *First sexual intercourse*
Atribut ini menunjukkan usia pertama kali pasien ketika melakukan hubungan seksual. Nilai pada atribut ini memiliki rentang antara 10 – 32 tahun.

d. *Number of pregnancies*
Merupakan jumlah kehamilan yang dialami pasien. Atribut ini memiliki rentang nilai 0 – 11 kali.

e. *Smokes*
Atribut ini menunjukkan apakah pasien merokok / tidak merokok. Diisi dengan nilai 1 / 0.

f. *Smokes (years)*
Meunjukkan berapa lama (dalam tahun) pasien yang memiliki kebiasaan merokok. Atribut ini memiliki rentang nilai 0 – 37 tahun.

g. *Smokes (packs/years)*
   Atribut ini menunjukkan berapa bungkus rokok yang dikonsumsi per tahunnya oleh pasien yang memiliki kebiasaan merokok. Memiliki rentang nilai antara 0 – 37 bungkus per tahun.
h. *Hormonal Contraceptives*
   Merupakan atribut yang menunjukkan apakah pasien menggunakan kontrasepsi hormonal atau tidak. Diisi dengan nilai 0 / 1.
i. *Hormonal Contraceptives (years)*
   Menunjukkan berapa lama (dalam tahun) pasien yang menggunakan kontrasepsi hormonal. Memiliki rentang nilai 0 – 22 tahun.
j. *IUD*
   Merupakan atribut yang menunjukkan apakah pasien menggunakan alat kontrasepsi *Intrauterine Device* (IUD) atau tidak. Diisi dengan nilai 0 / 1.
k. *IUD (years)*
   Menunjukkan berapa lama (dalam tahun) pasien yang menggunakan alat kontrasepsi IUD. Memiliki rentang nilai 0 – 19 tahun.
l. *STDs*
   Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin, atau *sexually transmitted diseases* (STDs). Diisi dengan nilai 0 / 1.
m. *STDs (number)*
   Menunjukkan jumlah penyakit kelamin yang pernah diderita oleh pasien. Memiliki rentang nilai 0 – 4 penyakit.
n. *STDs : condylomatosis*
   Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular *condylomatosis*. Diisi dengan nilai 0 /1.
o. *STDs : cervical condylomatosis*
   Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular *cervical condylomatosis.* Diisi dengan nilai 0 / 1.

p. *STDs : vaginal condylomatosis*
Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular *vaginal condylomatosis.* Diisi dengan nilai 0 / 1.

q. *STDs : vulvo-perineal condylomatosis*
Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular *vulvo-pernineal condylomatosis.* Diisi dengan nilai 0 /1.

r. *STDs : syphilis*
Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular syphilis. Diisi dengann nilai 0 / 1.

s. *STDs : pelvic inflammatory disease*
Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular *pelvic inflammatory disesase*. Diisi dengna nilai 0 / 1.

t. *STDs : genital herpes*
Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular *genital herpes*. Diisi dengan nilai 0 / 1.

u. *STDs : molluscum contagiosum*
Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular *molluscum contagiosum*. Diisi dengan nilai 0 / 1.

v. *STDs : AIDS*
Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular *AIDS*. Diisi dengan nilai 0 / 1.

w. *STDs : HIV*
Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular *HIV*. Diisi dengan nilai 0 / 1.

x. *STDs : Hepatitis B*
Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular *hepatitis B.* Diisi dengan nilai 0 / 1.

y. *STDs : HPV*
Merupakan atribut yang menunjukkan apakah pasien memiliki riwayat penyakit kelamin menular *HPV*. Diisi dengan nilai 0 / 1.

z. *STDs : Number of diagnosis*
Menunjukkan jumlah penyakit kelamin yang pernah didiagnosis pasien. Memiliki rentang nilai 0 – 3.

aa. *Biopsy*
Merupakan atribut target untuk penelitian. Menunjukkan hasil tes *biopsy* yang telah dilakukan oleh pasien. Diisi dengan nilai 0 / 1.

4.3 Pembersihan Data

Pembersihan data dilakukan dengan manual, yaitu menghilangkan beberapa baris data yang terdapat *missing value*. Data yang tidak lengkap akan berpengaruh pada hasil prediksi. Baris data yang tidak lengkap dihilangkan, dan data akhir yang digunakan menjadi bersih dari *missing value*. Data akhir yang siap digunakan untuk prediksi sebanyak 668 dari 858 data yang didapatkan.

Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD
18	4	15	1	0	0	0	0	0	0
15	1	14	1	0	0	0	0	0	0
52	5	16	4	1	37	37	1	3	0
46	3	21	4	0	0	0	1	15	0
42	3	23	2	0	0	0	0	0	0
51	3	17	6	1	34	3.4	0	0	1
26	1	26	3	0	0	0	1	2	1
45	1	20	5	0	0	0	0	0	0
44	3	26	4	0	0	0	1	2	0
27	1	17	3	0	0	0	1	8	0
45	4	14	6	0	0	0	1	10	1
44	2	25	2	0	0	0	1	5	0

Gambar 4.2. Tampilan Data

4.4 Prediksi Data

Proses prediksi data dimulai dari proses training untuk mencari pemisah antara dua kelas. Penelitian ini menggunakan *K-Fold Cross Validation* dengan 10 fold untuk melakukan proses training. Prediksi data akan menggunakan SVM kernel *linear* dan RBF.

4.4.1 Penentuan Parameter Kernel

Untuk menentukan parameter kernel, penulis akan melakukan prediksi dengan nilai parameter C dan γ yang berbeda. Nilai parameter C dan γ tersebut digunakan pada data training untuk mendapatkan nilai akurasi prediksi. Menentukan nilai parameter dapat dilakukan dengan cara manual dan otomatis. Salah satu metode yang dapat digunakan untuk menentukan parameter secara otomatis adalah *GridSearch*.[6] Namun pada penelitian tugas akhir ini, akan dilakukan penentuan parameter secara manual mulai dari 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, dan 10000.

4.4.2 *Resample Data*

Data yang digunakan untuk prediksi sebanyak 668 baris data, terdiri dari 623 label negatif kanker (0) dan 45 label positif kanker (1). Data yang ada tidak seimbang (*imbalanced*), sehingga dalam penelitian ini dilakukan *resampling* data. Data label negatif kanker akan di*downsampling* menjadi 100 baris data, dan label positif kanker akan di*umpsampling* menjadi 100 baris.

[6] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

BAB V
IMPLEMENTASI

Pada bab ini dijelaskan proses implementasi dalam pembuatan model prediksi. Implementasi membahas tentang perangkat apa saja yang digunakan, dan prediksi yang dilakukan dengan algoritma SVM.

5.1 Lingkungan Implementasi

Dalam penelitian prediksi diagnosa kanker serviks berdasarkan informasi demografik, kebiasaan, dan histori rekam medis menggunakan algoritma SVM, dibutuhkan perangkat – perangkat yang dapat mendukung proses pada tahapan penelitian. Perangkat – perangkat yang dibutuhkan meliputi perangkat keras dan perangkat lunak adalah sebagai berikut :

Perangkat Keras

- Komputer / laptop pribadi

Perangkat Lunak

- Sistem Operasi : Windows
- Bahasa Pemrograman : Python
- Tools : Anaconda Jupyter
  Ms. Excel
  Google Chrome

5.2 Implementasi Prediksi dengan SVM

Implementasi prediksi SVM adalah implementasi model prediksi data diagnosis kanker serviks yang telah dilakukan dengan menggunakan algoritma SVM. Implementasi dilakukan dengan bahasa pemrograman python dan menggunakan tools Anaconda Jupyter.

5.2.1 Import Library Python

Untuk melakukan prediksi menggunakan python, diperlukan beberapa *library* yang sesuai. *Library* yang digunakan adalah sebagai berikut :

```
import numpy as np
import pandas as pd

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.metrics import precision_recall_fscore_support
from sklearn.metrics import confusion_matrix

from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold

from sklearn.utils import resample
```

Kode 5.1. *Library* Python yang Digunakan

Berikut adalah penjelasan singkat mengenai *library* yang digunakan :

Tabel 5.1. *Library* Python yang Digunakan

Library	***Utility***
numpy	Menyediakan objek matematika yang mempermudah perhitungan. Objek yang disediakan adalah *array* dalam bentuk *matrix*.
pandas	Sebuah library open source yang menyediakan struktur data dan analisis data yang mudah digunakan dan memiliki kinerja tinggi untuk bahasa pemrograman Python. *Pandas* memiliki struktur data yang diperlukan untuk membersihkan data mentah ke dalam sebuah bentuk yang cocok untuk analisis.

Berikut macam – macam *function* yang digunakan dalam proses prediksi data :

Tabel 5.2. *Library* dan Fungsi untuk Prediksi

Library	***Function***	***Utility***
statistics	*mean*	Menghitung rata – rata *mean* dari data
sklearn	*svm*	*Library sklearn* untuk metode *Support Vector Machine*
sklearn .metrics	*classification_ report*	Membuat laporan yang menunjukkan hasil klasifikasi
	accuracy_score	Menghitung nilai akurasi dari klasifiksi
	mean_absolute_ error	Menghitung nilai eror dari klasifikasi
	precision_recall_ fscore_support	Menghitung nilai *precision, recall,* dan *f-measure*.
	confusion_matrix	Menghitung *confusion matrix* untuk mengevaluasi akurasi klasifikasi
sklearn. model_ selections	*train_test_split*	Untuk memisahkan data *array* atau *matrix* kedalam subset *train* dan *test* secara acak
	KFold	Menyediakan indeks *train / test* untuk membagi data kedalam lipatan (*fold*) ksecara berurutan
sklearn. utils	*resample*	Melakukan *resample* dengan konsisten

5.2.2 Load dan Read Data

Untuk memasukkan data berformat .csv, menggunakan library *pandas*. Kemudian data .csv diubah menjadi *matrix*, dan mendefinisikan data train dan data target berdasarkan indeks kolom dari array. data_train adalah 26 atribut yang termasuk faktor risiko diagnosa kanker serviks. data_target adalah satu atribut yaitu hasil tes *biopsy* pasien.

```
data = pd.read_csv('27var.csv', delimiter=';')

data_matrix = data.as_matrix()

data_train = np.array(data_matrix[:, 0:26])
data_target = np.array(data_matrix[:,26])
```

Kode 5.2. *Load* dan *Read Data*

5.2.3 *Resample Data*

Percobaan *resample* data dilakukan setelah melakukan percobaan dengan data yang memiliki 668 baris data. Data mayoritas untuk label negatif kanker (0) akan *downsampling* menjadi 100 baris data, dan untuk label positif kanker (1) akan *upsampling* menjadi 100 baris data. Kedua label data tersebut akan di*sampling* secara random, sehingga data akhir yang digunakan pada tahapan *resample data* merupakan dataset baru, berbeda dengan data yang di *load* dari file .csv sebelumnya.

Data dilakukan *resample* secara terpisah, seperti pada tampilan Kode 5.3. Setelah mendapatkan *data_negatif_downsampled* dan *data_positif_upsampled*, kedua data tersebut digabung dengan fungsi *concat*. Namun kelemahan dari fungsi *concat* ini, data tidak bisa menjadi random. Karena label negatif kanker ditulis terlebih dahulu, maka urutan baris data 1 – 100 adalah label negatif kanker, dan 101 – 200 adalah label positif kanker.

```
data_negatif = data[data.Biopsy==0]
data_positif = data[data.Biopsy==1]

data_negatif_downsampled = resample(data_negatif,
                                    replace=False,
                                    n_samples=100,
                                    random_state=123)
data_positif_upsampled = resample(data_positif,
                                  replace=True,
                                  n_samples=100,
                                  random_state=123)

data_sampled = pd.concat([data_negatif_downsampled,
                          data_positif_upsampled])
```

Kode 5.3. *Resample* Data

5.2.4 Menjalakan Model Prediksi

Pada penelitian ini menggunakan metode *Cross Validation* dengan jumlah *fold* = 10. Prediksi dilakukan dengan SVM kernel *linear* dan RBF. Nilai C dan γ yang digunakan adalah 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, dan 1000. Jumlah percobaan perhitungan dari nilai C dan γ adalah 100 percobaan pada 668 data, dan 100 percobaan lagi pada 200 data *resample.*

```
kf = KFold(n_splits=10)

C = 1.0
gamma = 10.0
kernel = 'rbf'
```

Kode 5.4. *Cross Validation* dan Penentuan Parameter

Untuk melakukan proses *training*, menentukan indeks yang akan dijadikan data *training* dan data *testing* berdasarkan 10 *fold*. Berikut adalah kode dari proses *training* dan *testing*.

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

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

    classifier = svm.SVC(kernel=kernel, C=C, gamma=gamma).fit(X_trainSet, y_trainSet)

    prediction = classifier.predict(X_testSet)
```

Kode 5.5. *Training* dan *Testing*

5.2.5 Melihat Performa

Tahapan selanjutnya yaitu melihat hasil performa prediksi yang telah dilakukan. Pengukuran performa prediksi ditulis seperti kode berikut :

```
precision,recall,fscore,Null_Value=precision_recall_fscore_support(y_testSet,
                                                                   prediction,
                                                                   average='macro')
```

Kode 5.6. Menghitung Pengukuran Performa

Sebelumnya kita harus membuat array untuk menyimpan label aktual dan hasil prediksi dari seluruh fold, serta array untuk menyimpan hasil pengukuran performa dari setiap fold.

```
final_test_labels = []
final_prediction = []

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

Kode 5.7. Menyimpan Label dan Hasil Pengukuran Performa

```
precision_val.append(precision)
recall_val.append(recall)
fscore_val.append(fscore)

final_test_labels.extend(y_testSet)
final_prediction.extend(prediction)
```

Kode 5.8. Menambahkan Nilai Hasil Pengukuran dan Label di Semua Fold

Kemudian untuk melihat hasil performa prediksi menggunakan *statistics mean* dan *accuracy score*. Jika ingin melihat nilai masing – masing *precision*, *recall*, dan *f-measure* pada masing – masing label, dapat menggunakan *classification report*. Serta *confusion matrix* untuk menampilkan jumlah antara data aktual dan hasil prediksi.

```
print("avg presisi :", mean(precision_val))
print("avg recall :", mean(recall_val))
print("avg fscore:", mean(fscore_val))

print("akurasi :", accuracy_score(final_test_labels, final_prediction))

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

Kode 5.9. Menampilkan Hasil Performa Prediksi

Halaman ini sengaja dikosongkan.

BAB VI
HASIL DAN PEMBAHASAN

Pada bab ini akan dijelaskan hasil serta analisis terhadap hasil yang diperoleh dari proses impelementasi sebelumnya.

6.1 Hasil Prediksi SVM *Linear* 668 Data

Tabel 6.1. Hasil Prediksi SVM *Linear* 668 Data

C	Precision	Recall	F-Measure	Akurasi
0.0001	**0.4662**	**0.5**	**0.4824**	**0.9326**
0.001	**0.4662**	**0.5**	**0.4824**	**0.9326**
0.01	**0.4662**	**0.5**	**0.4824**	**0.9326**
0.1	**0.4662**	**0.5**	**0.4824**	**0.9326**
1	**0.4662**	**0.5**	**0.4824**	**0.9326**
10	0.4661	0.4975	0.4812	0.9281
100	0.4661	0.4975	0.4812	0.9281
1000	0.4661	0.4968	0.4808	0.9266
10000	0.5167	0.5035	0.4946	0.9251

Berdasarkan percobaan yang dilakukan menggunakan SVM *Linear*, dapat diketahui nilai akurasi tertinggi didapatkan pada **C = 0.0001** hingga **C = 1**dengan akurasi sebesar **0.9326 / 93.26 %**. Percobaan dengan C mulai dari 0.0001 hingga 1 menghasilkan nilai *Precision* sebesar 0.4662, *Recall* sebesar 0.5,dan *F-Measure* sebesar 0.4824. Berikut adalah *confusion matrix* pada nilai C dari 0.0001 hingga 1 :

Tabel 6.2. *Confusion Matrix* Kernel *Linear* 668 Data pada Akurasi Tertinggi

		Prediksi	
		0	**1**
Data Testing (Aktual)	**0**	623	0
	1	45	0

Pada Tabel 6.2, dapat dilihat bahwa nilai *True Negative* label 0 adalah 623, dan nilai *False Positive* label 0 adalah 0. Hal tersebut menunjukkan bahwa sistem dapat melakukan prediksi secara akurat 100 % pada label 0 (negatif kanker), karena nilai pada *True Positive* sama dengan jumlah data aktual pada label 0. Kemudian untuk nilai *False Negative* label 1 adalah 45, dan nilai *True Positive* label 1 adalah 0. Hal tersebut menunjukkan bahwa sistem tidak dapat melakukan prediksi pada label 1 (positif kanker), karena semua label diprediksi menjadi label 0.

6.2 Hasil Prediksi SVM RBF 668 Data

Tabel 6.3. Hasil Prediksi SVM RBF 668 Data

C	γ	Precision	Recall	F-Measure	Akurasi
0.00001	0.0001	0.4662	0.5	0.4824	0.9326
	0.001	0.4662	0.5	0.4824	0.9326
	0.01	0.4662	0.5	0.4824	0.9326
	0.1	0.4662	0.5	0.4824	0.9326
	1	0.4662	0.5	0.4824	0.9326
	10	0.4662	0.5	0.4824	0.9326
	100	0.4662	0.5	0.4824	0.9326
	1000	0.4662	0.5	0.4824	0.9326
	10000	0.4662	0.5	0.4824	0.9326

C	γ	Precision	Recall	F-Measure	Akurasi
0.0001	0.0001	0.4662	0.5	0.4824	0.9326
	0.001	0.4662	0.5	0.4824	0.9326
	0.01	0.4662	0.5	0.4824	0.9326
	0.1	0.4662	0.5	0.4824	0.9326
	1	0.4662	0.5	0.4824	0.9326
	10	0.4662	0.5	0.4824	0.9326
	100	0.4662	0.5	0.4824	0.9326
	1000	0.4662	0.5	0.4824	0.9326
	10000	0.4662	0.5	0.4824	0.9326
0.001	0.0001	0.4662	0.5	0.4824	0.9326
	0.001	0.4662	0.5	0.4824	0.9326
	0.01	0.4662	0.5	0.4824	0.9326
	0.1	0.4662	0.5	0.4824	0.9326
	1	0.4662	0.5	0.4824	0.9326
	10	0.4662	0.5	0.4824	0.9326
	100	0.4662	0.5	0.4824	0.9326
	1000	0.4662	0.5	0.4824	0.9326
	10000	0.4662	0.5	0.4824	0.9326
0.1	0.0001	0.4662	0.5	0.4824	0.9326
	0.001	0.4662	0.5	0.4824	0.9326
	0.01	0.4662	0.5	0.4824	0.9326
	0.1	0.4662	0.5	0.4824	0.9326
	1	0.4662	0.5	0.4824	0.9326
	10	0.4662	0.5	0.4824	0.9326
	100	0.4662	0.5	0.4824	0.9326

C	γ	Precision	Recall	F-Measure	Akurasi
	1000	0.4662	0.5	0.4824	0.9326
	10000	0.4662	0.5	0.4824	0.9326
1	0.0001	0.4662	0.5	0.4824	0.9326
	0.001	0.4662	0.5	0.4824	0.9326
	0.01	0.4662	0.5	0.4824	0.9326
	0.1	0.4662	0.5	0.4824	0.9326
	1	0.4662	0.5	0.4824	0.9326
	10	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	100	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	1000	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	10000	**0.5677**	**0.5291**	**0.5281**	**0.9356**
10	0.0001	0.4662	0.5	0.4824	0.9326
	0.001	0.4662	0.5	0.4824	0.9326
	0.01	0.5422	0.5082	0.5036	0.9251
	0.1	0.5166	0.5068	0.4999	0.9236
	1	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	10	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	100	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	1000	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	10000	**0.5677**	**0.5291**	**0.5281**	**0.9356**
100	0.0001	0.4662	0.5	0.4824	0.9326
	0.001	0.4661	0.4983	0.4816	0.9296
	0.01	0.6088	0.5406	0.5407	0.9056
	0.1	0.4933	0.5130	0.5026	0.9056
	1	**0.5677**	**0.5291**	**0.5281**	**0.9356**

C	γ	Precision	Recall	F-Measure	Akurasi
	10	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	100	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	1000	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	10000	**0.5677**	**0.5291**	**0.5281**	**0.9356**
1000	0.0001	0.4661	0.4991	0.4819	0.9311
	0.001	0.4912	0.4958	0.4869	0.9161
	0.01	0.5456	0.5249	0.5270	0.8862
	0.1	0.4933	0.5130	0.5026	0.9056
	1	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	10	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	100	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	1000	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	10000	**0.5677**	**0.5291**	**0.5281**	**0.9356**
10000	0.0001	0.4661	0.4983	0.4816	0.9296
	0.001	0.5435	0.4997	0.5036	0.8952
	0.01	0.5246	0.5268	0.5185	0.8413
	0.1	0.4933	0.5130	0.5026	0.9056
	1	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	10	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	100	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	1000	**0.5677**	**0.5291**	**0.5281**	**0.9356**
	10000	**0.5677**	**0.5291**	**0.5281**	**0.9356**

Berdasarkan percobaan yang dilakukan menggunakan SVM RBF, dapat diketahui nilai akurasi tertinggi didapatkan pada **C = 1** hingga **C = 10000** dan $\boldsymbol{\gamma}$ **= 1** hingga $\boldsymbol{\gamma = 10000}$ dengan akurasi sebesar **0.9356 / 93.56 %**. Percobaan tersebut menghasilkan nilai *Precision* sebesar 0.5677, *Recall* sebesar 0.5291, dan *F-Measure* sebesar 0.5281. Berikut adalah *confusion matrix* yang didapatkan :

Tabel 6.4. *Confusion Matrix* pada Kernel RBF 668 Data pada Akurasi Tertinggi

		Prediksi	
		0	**1**
Data Testing (Aktual)	**0**	623	0
	1	45	0

Pada Tabel 6.4, dapat dilihat bahwa nilai *True Negative* label 0 adalah 623, dan nilai *False Positive* label 0 adalah 0. Hal tersebut menunjukkan bahwa sistem dapat melakukan prediksi secara akurat 100 % pada label 0 (negatif kanker), karena nilai pada *True Positive* sama dengan jumlah data aktual pada label 0. Kemudian untuk nilai *False Negative* label 1 adalah 45, dan nilai *True Positive* label 1 adalah 0. Hal tersebut menunjukkan bahwa sistem tidak dapat melakukan prediksi pada label 1 (positif kanker), karena semua label diprediksi menjadi label 0.

6.3 Hasil Prediksi SVM *Linear* 200 Data

Tabel 6.5. Hasil Prediksi SVM *Linear* 200 Data

C	Precision	Recall	F-Measure	Akurasi
0.0001	0	0	0	0
0.001	0.25	0.07	0.1083	0.14
0.01	0.5	0.1975	0.2785	0.395
0.1	**0.5**	**0.255**	**0.3344**	**0.51**
1	0.5	0.2525	0.3319	0.505
10	0.5	0.2425	0.3230	0.485
100	0.5	0.245	0.3259	0.49
1000	0.5	0.245	0.3255	0.49
10000	0.5	0.245	0.3232	0.49

Berdasarkan percobaan dari **200 data *sampling*** menggunakan SVM *Linear*, dapat diketahui nilai akurasi tertinggi didapatkan pada **C = 0.1** dengan akurasi sebesar **0.51 / 51 %.** Percobaan dengan C = 1 menghasilkan nilai *Precision* sebesar 0.5, *Recall* sebesar 0.255, dan *F-Measure* sebesar 0.3344. Berikut adalah *confusion matrix* pada nilai C = 0.1.

Tabel 6.6. *Confusion Matrix* pada Kernel *Linear* 200 Data Sample pada Akurasi Tertinggi

		Prediksi	
		0	**1**
Data Testing (Aktual)	**0**	58	42
	1	56	44

Pada Tabel 6.6, dapat dilihat bahwa nilai *True Negative* label 0 adalah 58, dan nilai *False Positive* label 0 adalah 42. Hal tersebut menunjukkan bahwa sistem dapat melakukan prediksi secara akurat hanya 50 % pada label 0 (negatif kanker). Kemudian untuk nilai *False Negative* label 1 adalah 56, dan nilai *True Positive* label 1 adalah 44. Hal tersebut menunjukkan bahwa sistem dapat melakukan prediksi secara akurat hanya 44% pada label 1 (positif kanker).

6.4 Hasil Prediksi SVM RBF 200 Data

Tabel 6.7. Hasil Prediksi SVM *Linear* 200 Data

C	γ	Precision	Recall	F-Measure	Akurasi
0.0001	0.0001	0	0	0	0
	0.001	0	0	0	0
	0.01	0	0	0	0
	0.1	0	0	0	0
	1	0	0	0	0
	10	0	0	0	0
	100	0	0	0	0
	1000	0	0	0	0
	10000	0	0	0	0
0.001	0.0001	0	0	0	0
	0.001	0	0	0	0
	0.01	0	0	0	0
	0.1	0	0	0	0
	1	0	0	0	0
	10	0	0	0	0

C	γ	Precision	Recall	F-Measure	Akurasi
	100	0	0	0	0
	1000	0	0	0	0
	10000	0	0	0	0
0.01	0.0001	0	0	0	0
	0.001	0	0	0	0
	0.01	0	0	0	0
	0.1	0	0	0	0
	1	0	0	0	0
	10	0	0	0	0
	100	0	0	0	0
	1000	0	0	0	0
	10000	0	0	0	0
0.1	0.0001	0	0	0	0
	0.001	0	0	0	0
	0.01	0	0	0	0
	0.1	0	0	0	0
	1	0	0	0	0
	10	0	0	0	0
	100	0	0	0	0
	1000	0	0	0	0
	10000	0	0	0	0
1	0.0001	0	0	0	0
	0.001	0.35	0.09	0.1390	0.18
	0.01	0.5	0.215	0.2939	0.43
	0.1	0.5	0.375	0.4252	0.75

C	γ	Precision	Recall	F-Measure	Akurasi
	1	0.7	0.6575	0.6764	0.915
	10	**0.75**	**0.71**	**0.7277**	**0.92**
	100	**0.75**	**0.71**	**0.7277**	**0.92**
	1000	**0.75**	**0.71**	**0.7277**	**0.92**
	10000	**0.75**	**0.71**	**0.7277**	**0.92**
10	0.0001	0.3	0.0775	0.1183	0.155
	0.001	0.5	0.19	0.2701	0.38
	0.01	0.5	0.3025	0.3743	0.605
	0.1	0.5	0.4175	0.4543	0.835
	1	0.7	0.6575	0.6764	0.915
	10	**0.75**	**0.71**	**0.7277**	**0.92**
	100	**0.75**	**0.71**	**0.7277**	**0.92**
	1000	**0.75**	**0.71**	**0.7277**	**0.92**
	10000	**0.75**	**0.71**	**0.7277**	**0.92**
100	0.0001	0.5	0.2075	0.2859	0.415
	0.001	0.5	0.25	0.3273	0.5
	0.01	0.5	0.3475	0.4065	0.695
	0.1	0.5	0.4175	0.4543	0.835
	1	0.7	0.6575	0.6764	0.915
	10	**0.75**	**0.71**	**0.7277**	**0.92**
	100	**0.75**	**0.71**	**0.7277**	**0.92**
	1000	**0.75**	**0.71**	**0.7277**	**0.92**
	10000	**0.75**	**0.71**	**0.7277**	**0.92**
1000	0.0001	0.5	0.24	0.3187	0.48
	0.001	0.5	0.3025	0.3736	0.605

C	γ	Precision	Recall	F-Measure	Akurasi
	0.01	0.5	0.3775	0.4278	0.755
	0.1	0.5	0.4175	0.4543	0.835
	1	0.7	0.6575	0.6764	0.915
	10	**0.75**	**0.71**	**0.7277**	**0.92**
	100	**0.75**	**0.71**	**0.7277**	**0.92**
	1000	**0.75**	**0.71**	**0.7277**	**0.92**
	10000	**0.75**	**0.71**	**0.7277**	**0.92**
10000	0.0001	0.5	0.29	0.3627	0.58
	0.001	0.5	0.33	0.3933	0.66
	0.01	0.5	0.38	0.4280	0.76
	0.1	0.5	0.4175	0.4543	0.835
	1	0.7	0.6575	0.6764	0.915
	10	**0.75**	**0.71**	**0.7277**	**0.92**
	100	**0.75**	**0.71**	**0.7277**	**0.92**
	1000	**0.75**	**0.71**	**0.7277**	**0.92**
	10000	**0.75**	**0.71**	**0.7277**	**0.92**

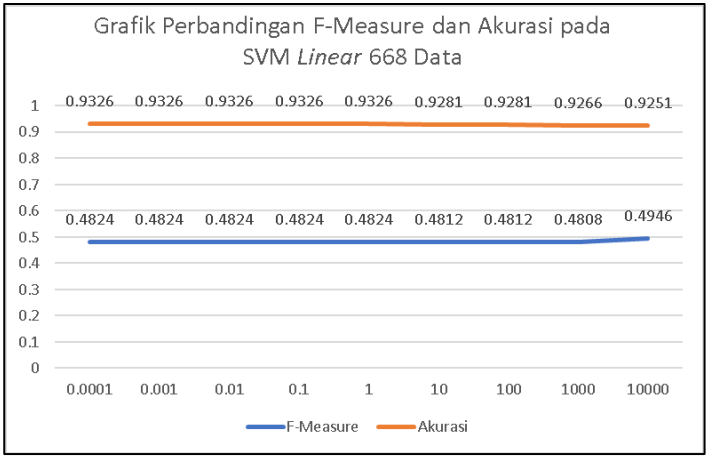
Berdasarkan percobaan dari **200 data *sampling*** menggunakan SVM RBF, dapat diketahui nilai akurasi tertinggi didapatkan pada **C = 1** hingga **C = 10000** dan $\boldsymbol{\gamma}$ **= 10** hingga $\boldsymbol{\gamma = 10000}$ dengan akurasi sebesar **0.92 / 92 %.** Percobaan tersebut menghasilkan nilai *Precision* sebesar 0.75, *Recall* sebesar 0.71, dan *F-Measure* sebesar 0.7277. Berikut adalah *confusion matrix* yang didapatkan :

Tabel 6.8. *Confusion Matrix* pada Kernel *Linear* 200 Data Sample pada Akurasi Tertinggi

		Prediksi	
		0	**1**
Data Testing (Aktual)	**0**	100	0
	1	16	84

Pada Tabel 6.8 dapat dilihat bahwa nilai *True Negative* label 0 adalah 100, dan nilai *False Positive* label 0 adalah 0. Hal tersebut menunjukkan bahwa sistem dapat melakukan prediksi secara akurat 100 % pada label 0 (negatif kanker), karena nilai pada *True Positive* sama dengan jumlah data aktual *resampled* pada label 0. Kemudian untuk nilai *False Negative* label 1 adalah 16, dan nilai *True Positive* label 1 adalah 84. Hal tersebut menunjukkan bahwa sistem dapat melakukan prediksi pada label 1 secara akurat 84 %.

6.5 Pembahasan Prediksi SVM 668 Data

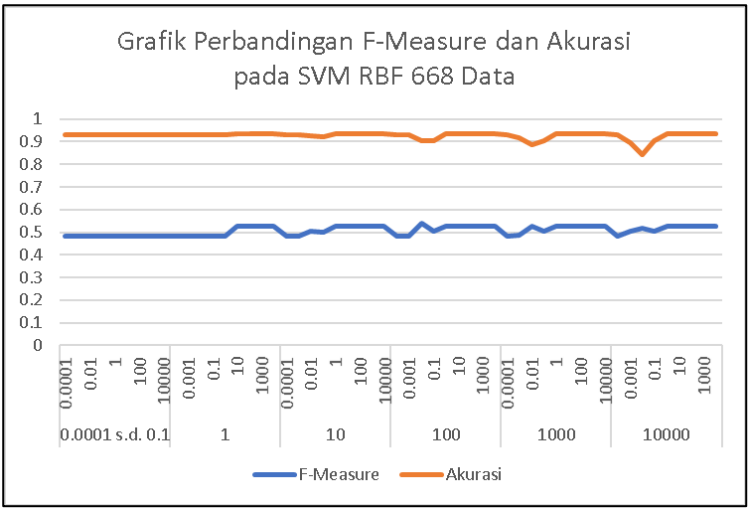


Gambar 6.1. Grafik Perbandingan *F-Measure* dan Akurasi pada SVM *Linear* 668 Data

Pada hasil prediksi dari 668 data menggunakan SVM kernel *linear*,dapat dilihat dalam Tabel 6.1 bahwa mulai dari parameter C = 0.0001 hingga C = 1 memiliki nilai *Precision, Recall, F-Measure,* dan akurasi yang sama. Parameter C = 10 hingga C = 1000 juga memiliki nilai *Precision, Recall, F-Measure,* dan akurasi yang sama. Dari hasil pengukuran yang didapatkan dengan SVM *linear* 668 data, semua parameter C yang dicoba menghasilkan nilai akurasi > 90 %.

Pada hasil prediksi dari 668 data menggunakan SVM kernel RBF, dapat dilihat dalam Tabel 6.3 bahwa mulaidari parameter C = 0.00001 hinga C = 0.1, dan nilai γ = 0.0001 hingga γ = 10000, memiliki nilai *Precision, Recall, F-Measure* dan akurasi yang sama. Bahkan hasilnya sama dengan hasil prediksi menggunakan kernel *linear* pada C = 0.0001 hingga C = 1.

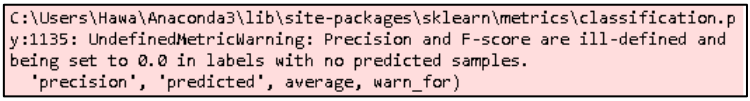
Hasil akurasi terbaik didapatkan dari parameter C = 1 hingga C = 10000 dan γ = 1 hingga γ = 10000. Meskipun dengan kernel RBF memiliki nilai akurasi sedikit lebih tinggi (93.56%) dibandingkan dengan ketika menggunakan kernel *linear* (93.26%), hasil perhitungan nilai *Precision*, *Recall*, dan *F-Measure* lebih rendah dari nilai akurasinya. Dari hasil pengukuran yang didapatkan dengan SVM RBF 668 data, semua parameter C dan γ dan yang dicoba menghasilkan nilai akurasi > 80 %.



Gambar 6.2. Grafik Perbandingan *F-Measure* dan Akurasi pada SVM RBF 668 Data

Hasil perbandingan antara nilai *F-measure* yang memiliki nilai lebih rendah daripada nilai akurasi menunjukkan bahwa model yang dihasiilkan dapat memprediksi data dengan baik, namun tingkat ketepatan dan tingkat keberhasilan untuk memprediksinya rendah. Hal tersebut dapat terjadi karena pada hasil *confusion matrix* pada label 1 (positif kanker) yang tidak tepat prediksinya. Data aktual label 1 terdapat sebanyak 45 data, namun semuanya diprediksi ke dalam label 0 (negatif kanker).

Berikut adalah tampilan *warning* ketika menjalankan model prediksi :

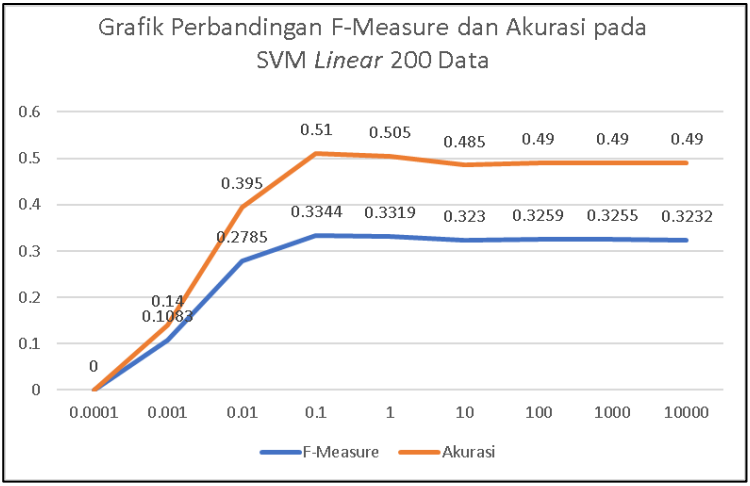


Gambar 6.3. *Warning* SVM Karena Tidak Menemukan Label Tujuan

Prediksi label 1 yang tidak tepat kemungkinan disebabkan pada saat melakukan training dengan *K-Fold Cross Validation*, terdapat data yang tidak memiliki label 1. Ketika model hasil data training tersebut dijalankan pada data testing,, data yang akan diprediksi tidak menemukan label tujuan, dan sistem akan secara otomatis memasukkan data tersebut ke dalam label 0.

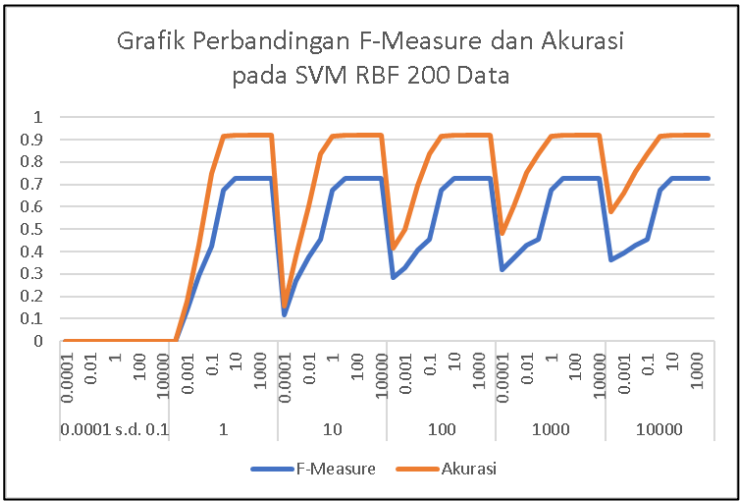
Akurasi prediksi pada 668 data memiliki nilai yang tinggi dan hampir mendekati 100 %. Hal tersebut dikarenakan model dapat memprediksi > 90 % data kedalam label yang benar, yaitu label 0 (negatif kanker). Label 0 berjumlah 623 dari 668 data yang ada, atau sebanyak 94 % dari jumlah data, sehingga nilai akurasi tertinggi didapatkan karena jumlah data yang tidak seimbang antara label 0 (negatif kanker) dan label 1 (positif kanker).

6.6 Pembahasan Prediksi SVM 200 Data



Gambar 6.4. Grafik Perbandingan *F-measure* dan Akurasi pada SVM *Linear* 200 Data

Pada hasil prediksi dari 200 data menggunakan SVM kernel *linear*,dapat dilihat bahwa dari semua parameter C yang dicoba menghasilkan akurasi yang rendah dengan rentang akurasi 0 – 50 %. Dari 9 percobaan dengan 200 data, nilai *recall* lebih rendah daripada *precision.* Hal ini menunjukkan bahwa model yang dihasilkan memiliki tingkat keberhasilan untuk menemukan suatu data lebih rendah daripada ketepatan prediksi. Hasil ini berbanding terbalik ketika 668 data dicoba dengan SVM *linear*, dimana semua parameter menghasilkan tingkat akurasi > 90 %. Dengan semua hasil akurasi yang rendah, dapat dikatakan bahwa prediksi dengan 200 data yang telah dilakukan *resample* kurang tepat jika menggunakan SVM kernel *linear*.



Gambar 6.5. Grafik Perbandingan *F-measure* dan Akurasi pada SVM RBF 200 Data

Pada hasil prediksi dari 200 data menggunakan SVM kernel RBF memiliki hasil yang tidak begitu berbeda dengan percobaan 668 data sebelumnya. Nilai akurasi tertinggi berada pada rentang C = 10 hingga 10000, dan nilai γ = 10 hingga 10000. Percobaan 200 data dengan SVM RBF inilah yang memiliki nilai akurasi, *precision*, *recall*, dan *f-measure* paling tinggi diantara tiga percobaan sebelumnya.

Pada percobaan SVM RBF 668 data, dari parameter γ 0.0001 – 10000 tidak terjadi perubahan yang begitu besar dalam tingkat akurasi. Sedangkan pada percobaan SVM RBF 200 data, nilai γ < 1 pada semua parameter C yang dicoba akan menghasilkan akurasi yang rendah.

Nilai akurasi pada 200 data (92 %) sedikit lebih rendah daripada 668 data (93.56%), namun memiliki nilai *precision*, *recall*, dan *f-measure* yang lebih tinggi. Dengan nilai *precision*, *recall*, dan *f-measure* > 70 %, dapat dikatakan bahwa model prediksi yang dihasilkan dengan SVM RBF menggunakan 200 data memiliki tingkat akurasi sangat tinggi, tingkat ketepatan cukup baik, dan tingkat keberhasilan juga cukup baik.

6.7 Pembahasan Hasil Akhir Prediksi dengan Penelitian Sebelumnya

Pada penelitian tugas akhir ini, data yang digunakan dilakukan pembersihan data terlebih dahulu, sehingga atribut / variabel yang kosong (*missing value*) tidak digunakan dalam penelitian. Atribut / variabel yang kosong tersebut tidak diisi dengan nilai rata – rata dari atribut / variabel seperti pada penelitian [6] dan penelitian [2]. Penulis menghilangkan atribut yang memiliki *missing value* tersebut karena tidak ingin mengubah keaslian data yang telah didapatkan. Penelitian ini juga menghilangkan beberapa atribut yang banyak memiliki *missing value*. Jadi, data yang siap digunakan untuk prediksi adalah data murni dari jawaban 668 pasien rumah sakit.

Nilai akurasi terbaik (yang memiliki nilai *precision, recall,* dan *f-measure* terbaik) pada empat percobaan yang dilakukan pada prediksi diagnosa kanker serviks dengan SVM ini adalah sebesar 92 %. Nilai ini lebih rendah daripada hasil penelitian [2] yang memiliki nilai akurasi 97.26 % dengan menggunakan k-NN. Pada penelitian [2] tidak dilakukan *resample data*, sehingga data yang digunakan adalah data yang tidak seimbang label kelasnya.

Data tidak seimbang tentu akan mempengaruhi hasil dan performa klasifikasi / prediksi data yang dilakukan. Dengan *imbalanced data* atau data tidak seimbang, kencenderungan label kelas data menjadi tidak stabil, karena data lebih condong ke bagian data mayoritas. Hasil akurasi saja bukanlah cara terbaik jika ingin mengukur keberhasilan dari suatu klasifikasi / prediksi yang telah dilakukan. *Resample data* dilakukan untuk menghindari kecenderungan data yang condong ke bagian mayoritas. Data yang seimbang akan membuat nilai *precision, recall,* dan *f-measure* menjadi seimbang dengan nilai akurasi yang didapatkan.

BAB VII
KESIMPULAN DAN SARAN

Pada bab ini akan dibahas mengenai kesimpulan dari pengerjaan tugas akhir yang telah dilakukan dan saran yang dapat diberikan sebagai pengembangan penelitian yang lebih baik.

7.1 Kesimpulan

Berdasarkan pengerjaan tugas akhir berjudul "Prediksi Diagnosa Kanker Serviks berdasarkan Informasi Demografi, Kebiasaan, dan Rekam Medis Menggunakan Algoritma *Support Vector Machine*" dapat disimpulkan beberapa hal sebagai berikut :

1. Nilai akurasi tertinggi didapatkan pada percobaan prediksi dengan SVM *RBF* menggunakan 668 data, yaitu sebesar 93.56 %. Namun memiliki nilai *precision, recall,* dan *f-measure* yang lebih rendah.
2. Hasil pengukuran terbaik didapatkan pada percobaan prediksi dengan SVM RBF menggunakan 200 data, dengan nilai akurasi sebesar 92 %, dan semua nilai *precision*, *recall,* dan *f-measure* > 70 %.
3. Untuk mengevaluasi hasil pengukuran prediksi dengan data yang tidak seimbang, sebaiknya tidak dilihat dari hasil akurasi saja. Evaluasi hasil pengukuran juga dilihat melalui nilai *precision*, *recall,* dan *f-measure*.
4. *Resample data* untuk data yang tidak seimbang (*imbalanced data)* membuat hasil prediksi pada data yang tidak seimbang menjadi lebih baik dalam segi akurasi, *precision*, *recall,* dan *f-measure*.

7.2 Saran

Berdasarkan hasil pengerjaan tugas akhir yang telah dilakukan, beberapa saran yang dapat dipertimbangkan untuk penelitian selanjutnya antara lain:

1. Melakukan *resample data* dengan jumlah yang berbeda berdasarkan prosentase, untuk meningkatkan nilai pengukuran performa akurasi, *precision*, *recall,* dan *f-measure.*
2. Melakukan prediksi terhadap atribut target selain hasil tes *Biopsy*. Pada data asli terdapat tiga atribut target lainnya, yaitu hasil tes *Hinselmann, Schiller,* dan *Cytology,* agar kedepannya dapat dilakukan perbandingan dari masing – masing hasil prediksi diagnosa kanker serviks.
3. Melakukan klasifikasi / prediksi dengan data yang telah dibersihkan atribut *missing value* dan dilakukan *resample* menggunakan algoritma selain SVM.

DAFTAR PUSTAKA

[1] T. Praningki and I. Budi, "Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN," *Citec Journal,* vol. 4, no. 2, 2017.

[2] M. F. Unlersen and K. Sabanci, "Determining Cervical Cancer Possibility by Using Machine Learning Methods," *Research Gate,* 2017.

[3] G. R. Kumar, D. G. A. Ramachandra and K. Nagamani, "An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 4, no. 2, 2014.

[4] P. D. d. I. K. K. RI, 2015. [Online]. Available: http://www.depkes.go.id/resources/download/pusdatin/infodatin/infodatin-kanker.pdf.

[5] P. Ramachandran, N. Girija and T. Bhuvaneswari, "Early Detection and Prevention of Cancer using Data Mining Techniques," *International Jurnal of Computer Application,* vol. 97, no. 13.

[6] K. Fernandes, J. S. Cardoso and J. Fernandes, "Trasnfer Learning with Partial Obersvability Applied to Cervical Cancer Screening," *Iberian Conference on Pattern Recognition and Image Analysis,* 2017.

[7] M. R. Dwi and S. W. Purnami, "Klasifikasi Hasil Pap Smear Test sebagai Upaya Pencegahan Sekunder Penyakit Kanker Serviks di Rumah Sakit "X" Surabaya Menggunakan Piecewise Polynomial Smooth Support Vector Machine (PPSSVM)," *Jurnal Sains dan Seni ITS,* vol. 4, no. 1, 2015.

[8] Komite Penanggulangan Kanker Nasional, [Online]. Available: http://kanker.kemkes.go.id/guidelines/PPKServiks.pdf.

[9] S. Syatriani, "Faktor Risiko Kanker Serviks di Rumah Sakit Umum Pemerintah Dr. Wahidin Sudirohusodo Makassar, Sulawesi Selatan," *Jurnal Kesehatan Masyarakat Nasional*, vol. 5, no. 6, 2017.

[10] S. N. Sivanandam and S. Sumathi, Introduction to Data Mining and Its Applications, Springer, 2006.

[11] J. Brownlee, "8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset," 2015. [Online]. Available: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/.

[12] D. Kelleher, B. M. Namee and A. D'Arcy, Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies, London: England : MIT Press, 2015.

[13] D. Triasanti, "Struktur Pemrograman Python," 2005. [Online]. Available: http://andriyani.staff.gunadarma.ac.id/Downloads/files/4.

LAMPIRAN A

Lampiran ini mencantumkan hasil perhitungan performa masing – masing fold pada parameter yang menghasilkan nilai akurasi terbaik.

Lampiran A. 1. Performa masing – masing fold pada SVM linear 668 Data Akurasi Terbaik

```
precision, recall, fscore, support
(0.46268656716417911, 0.5, 0.48062015503875971, None)
(0.48507462686567165, 0.5, 0.49242424242424243, None)
(0.45522388059701491, 0.5, 0.4765625, None)
(0.45522388059701491, 0.5, 0.4765625, None)
(0.47761194029850745, 0.5, 0.48854961832061067, None)
(0.47761194029850745, 0.5, 0.48854961832061067, None)
(0.47014925373134331, 0.5, 0.48461538461538461, None)
(0.46268656716417911, 0.5, 0.48062015503875971, None)
(0.43181818181818182, 0.5, 0.46341463414634143, None)
(0.48484848484848486, 0.5, 0.49230769230769234, None)
```

Lampiran A. 2. Hasil *classification report* pada SVM linear 668 Data Akurasi Terbaik

	precision	recall	f1-score	support
0.0	0.93	1.00	0.97	623
1.0	0.00	0.00	0.00	45
avg / total	0.87	0.93	0.90	668

Lampiran A. 3. Performa masing – masing fold pada SVM RBF 668 Data Akurasi Terbaik

```
precision, recall, fscore, support
(0.46268656716417911, 0.5, 0.48062015503875971, None)
(0.48507462686567165, 0.5, 0.49242424242424243, None)
(0.45522388059701491, 0.5, 0.4765625, None)
(0.45522388059701491, 0.5, 0.4765625, None)
(0.98484848484848486, 0.66666666666666663, 0.74230769230769234, None)
(0.47761194029850745, 0.5, 0.48854961832061067, None)
(0.97727272727272729, 0.625, 0.68837209302325586, None)
(0.46268656716417911, 0.5, 0.48062015503875971, None)
(0.43181818181818182, 0.5, 0.46341463414634143, None)
(0.48484848484848486, 0.5, 0.49230769230769234, None)
```

Lampiran A. 4. Hasil *classification report* pada SVM RBF 668 Data Akurasi Terbaik

	precision	recall	f1-score	support
0.0	0.94	1.00	0.97	623
1.0	1.00	0.04	0.09	45
avg / total	0.94	0.94	0.91	668

Lampiran A. 5. Performa masing – masing fold pada SVM *linear* 200 Data Akurasi Terbaik

```
precision, recall, fscore, support
(0.5, 0.29999999999999999, 0.37499999999999994, None)
(0.5, 0.25, 0.33333333333333331, None)
(0.5, 0.32500000000000001, 0.39393939393939398, None)
(0.5, 0.27500000000000002, 0.35483870967741937, None)
(0.5, 0.29999999999999999, 0.37499999999999994, None)
(0.5, 0.17499999999999999, 0.25925925925925924, None)
(0.5, 0.29999999999999999, 0.37499999999999994, None)
(0.5, 0.17499999999999999, 0.25925925925925924, None)
(0.5, 0.20000000000000001, 0.28571428571428575, None)
(0.5, 0.25, 0.33333333333333331, None)
```

Lampiran A. 6. Hasil *classification report* pada SVM *linear* 200 Data Akurasi Terbaik

	precision	recall	f1-score	support
0.0	0.51	0.58	0.54	100
1.0	0.51	0.44	0.47	100
avg / total	0.51	0.51	0.51	200

Lampiran A. 7. Performa masing – masing fold pada SVM RBF 200 Data Akurasi Terbaik

```
precision, recall, fscore, support
(1.0, 1.0, 1.0, None)
(1.0, 1.0, 1.0, None)
(1.0, 1.0, 1.0, None)
(1.0, 1.0, 1.0, None)
(1.0, 1.0, 1.0, None)
(0.5, 0.375, 0.42857142857142855, None)
(0.5, 0.45000000000000001, 0.47368421052631582, None)
(0.5, 0.375, 0.42857142857142855, None)
(0.5, 0.42499999999999999, 0.45945945945945943, None)
(0.5, 0.47499999999999998, 0.48717948717948717, None)
```

Lampiran A. 8. Hasil *classification report* pada SVM RBF 200 Data Akurasi Terbaik

	precision	recall	f1-score	support
0.0	0.86	1.00	0.93	100
1.0	1.00	0.84	0.91	100
avg / total	0.93	0.92	0.92	200

BIODATA PENULIS

Penulis bernama Siti Hawa Aminah, lahir di Sidoarjo, Jawa Timur pada tanggal 16 Januari 1994. Merupakan anak ketiga dari tiga bersaudara. Penulis telah menempuh pendidikan formal di SD Negeri Entalsewu, SMP Negeri 1 Sidoarjo, dan SMA Negeri 1 Sidoarjo.

Pada tahun 2012, 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) 5212100054. Selama masa perkuliahan, penulis aktif di berbagai organisasi mulai dari staff departemen Sosial Masyarakat HMSI 2013/2014, staff departemen Keputrian KISI 2013/2014, hingga pernah menjabat sebagai Sekretaris departemen Sosial Masyarakat HMSI 2014/2015, dan Ketua departemen Keputrian KISI 2014/2015. Penulis juga aktif diberbagai kepanitiaan pada acara GERIGI ITS, LKMM pra TD fakultas dan LKMM TD jurusan, dan beberapa kegiatan HMSI. Selain itu, penulis juga pernah menjadi asisten dosen pada mata kuliah Perencanaan Sumber Daya Perusahaan pada tahun 2015/2016. Penulis mengambil bidang minat laboratorium Akuisisi Data dan Diseminasi Informasi (ADDI) untuk pengerjaan Tugas Akhir.

Apabila terdapat pertanyaan mengenai Tugas Akhir ini, penulis dapat dihubungi melalui e-mail siti.hawa1601@gmail.com