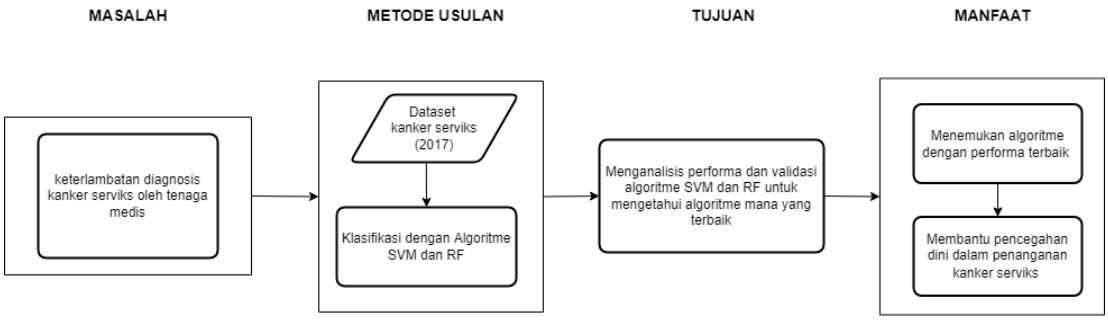
BAB III
METODOLOGI PENELITIAN

3.1 Kerangka Berpikir

Gambar 3.1 menunjukkan kerangka berpikir dari penelitian ini yang meliputi masalah, metode usulan, tujuan, dan manfaat. Masalah dari penelitian ini adalah kanker serviks menjadi penyebab kematian kedua wanita di dunia. Diagnosis yang dilakukan oleh tenaga medis juga mengalami keterlambatan sehingga kanker serviks tidak bisa diatasi sedini mungkin. Usulan yang diberikan pada penelitian ini adalah dengan melakukan pengklasifikasian dengan algoritma SVM dan RF terhadap *dataset* kanker serviks dengan tujuan menganalisis performa dan validasi dari algoritma SVM dan RF untuk mengetahui algoritma mana yang terbaik. Manfaat yang didapatkan adalah dengan menemukan algoritma dengan performa terbaik dapat membantu pencegahan dini dalam penanganan kanker serviks.



Gambar 3. 1 Kerangka Berpikir

3.2 Subjek dan Objek Penelitian

Subjek penelitian ini adalah algoritma klasifikasi yang memiliki akurasi terbaik pada penelitian sebelumnya, yaitu algoritma SVM dan RF. Dari kedua algoritma tersebut akan dicari kembali algoritma terbaik.

Objek penelitian ini adalah *dataset* mengenai risiko penyakit kanker serviks dengan label kelas adalah *Biopsy* dengan nilai 1 dan 0. Label kelas tersebut

memiliki arti bahwa berarti responden dengan dan tanpa terindikasi oleh penyakit kanker serviks. Kelas kanker serviks memiliki nilai 1 apabila responden terindikasi oleh kanker serviks dan bernilai 0 apabila responden tidak terindikasi oleh kanker serviks.

3.3 Teknik pengambilan data

Pengumpulan data diambil dari *dataset* bersifat public dari website UCI. *Dataset* dikumpulkan di "*Hospital Universitario de Caracas"* di Caracas, Venezuela yang terdiri dari informasi demografis, kebiasaan, dan catatan medis historis [68]. *Dataset* dapat diakses pada https://bit.ly/datasetrisikokankerserviks.

3.4 Alat dan bahan penelitian

Alat:

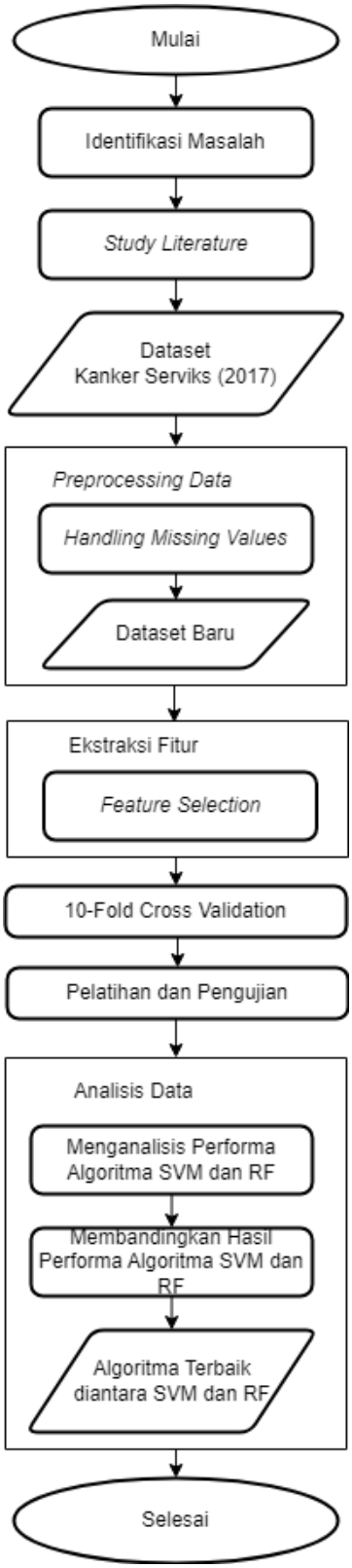
1. *Laptop Lenovo core i5*
2. *Google Colab, Visual Studio Code, Jupyter Notebook, WEKA, Rapidminer*

Bahan:

1. *Dataset* dari website UCI

3.5 Diagram alur penelitian

Alur penelitian ini adalah dengan menentukan topik yang kemudian dirumuskan masalah, tujuan, dan manfaatnya. Setelah itu, dilakukan pengambilan *dataset* dan akan dibagi menjadi data latih dan uji yang akan dilakukan latih menggunakan 10-Fold Cross Validation dengan algoritma *Random Forest* dan SVM. Langkah terakhir adalah melakukan analisis performa dari tiap algoritma. Gambar 3.2 menunjukkan diagram alur penelitian.



Gambar 3. 2 Alur Penelitian

3.6 *Dataset* Kanker Serviks (2017)

Dataset yang dipakai pada penelitian ini adalah *dataset* bersifat publik diambil dari website UCI seperti yang sudah dijelaskan. Adapun penjelasan mengenai *dataset* tersebut adalah sebagai berikut:

1. Data berjumlah 858.
2. Beberapa pasien memutuskan untuk tidak memberikan jawaban atas beberapa pertanyakan dikarenakan masalah privasi (*missing value*) [68].
3. Menurut [69] [70] target utama kelas dalam *dataset Cervical Cancer* adalah atribut *Biopsy*.

Pada Tabel 3.1 akan dijelaskan atribut yang terdapat pada *dataset*.

Tabel 3. 1 Deskripsi Atribut *Dataset*

Atribut	Deskripsi
Age	Merupakan umur pasien yang melakukan pemeriksaan.
Number of sexual partner	Jumlah partner seksual pada pasien yang melakukan pemeriksaan.
First sexual intercose	Merupakan umur dimana pasien berhubungan seksual pertama kali.
Num of pregnancies	Merupakan jumlah total kehamilan yang dialami pasien.
Smokes	Adalah pernyataan dimana pasien merupakan perokok ataukah tidak.
Smokes (years)	Merupakan umur dimana pasien merokok.
Smokes (packs/years)	Jumlah bungkus rokok yang dihabiskan pasien perokok tiap tahunnya.
Hormonal Contraceptives	Merupakan salah satu kontrasepsi dengan memanfaatkan hormon alami tubuh sehingga ovulasi tidak terjadi.
Hormonal Contraceptives (years)	Merupakan tahun dimana pasien menggunakan kontrasepsi jenis hormone.
IUD	Merupakan salah satu jenis alat kontrasepsi dengan bentuk spiral yang cara pemakaiannya dimasukkan ke dalam Rahim.
IUD (years)	Merupakan tahun dimana pasien menggunakan kontrasepsi jenis IUD.
STDs	Merupakan penyakit menular seksual yang diakibatkan oleh infeksi (bakteri, virus, atau penyakit) dan cara

Tabel 3. 1 Deskripsi Atribut *Dataset* (lanjutan)

Atribut	Deskripsi
	penularannya melalui hubungan seksual.
STDs (number)	Merupakan kode nomor dari jenis penyakit menular seksual.
STDs: condylomatosis	Salah satu jenis penyakit seksual menular yang diakibatkan oleh infeksi virus papillomavirus.
STDs: cervical condylomatosis	Merupakan kondilomatosis terjadi pada wanita yang terjadi pada rahim.
STDs: vaginal condylomatosis	Merupakan kondilomatosis terjadi pada wanita yang terjadi pada vagina.
STDs: vulvo-perineal condylomatosis	Merupakan kondilomatosis terjadi pada wanita yang terjadi pada vula perineum (kulit antara vagina dan anus).
STDs: syphilis	Salah satu penyakit menular seksual yang diakibatkan oleh bakteri, gejala diawali dengan luka tanpa perasaan sakit.
STDs: pelvic inflammatory disease	Salah satu penyakit menular seksual yang diakibatkan oleh bakteri yang menyebar mulai dari vagina ke uterus, tuba falopi, ataupun ovarium.
STDs: genital herpes	Salah satu penyakit menular seksual yang diakibatkan oleh virus herpes simplex, ditandai dengan luka atau nyeri pada alat kelamin.
STDs: molluscum contagiosum	Salah satu penyakit menular seksual yang diakibatkan oleh virus yang menyebabkan benjolan pada kulit, penyebarannya melalui tahapan kontak langsung pada orang yang sudah terinfeksi.
STDs: HIV	Salah satu penyakit menular seksual yang diakibatkan oleh virus HIV. Penyakit ini dapat menyebabkan penurunan imunitas, penyebarannya dapat melalui darah, air mani, ataupun cairan vagina.
STDs: AIDS	Salah satu penyakit menular seksual yang diakibatkan oleh virus HIV yang berkembang lebih lanjut.
STDs: Hepatitis B	Salah satu faktor STD yang berupa penyakit hati yang menular dan diakibatkan oleh virus hepatitis (HBV), tidak perlu penanganan khusus.
STDs: HPV	Salah satu penyakit menular seksual yang diakibatkan oleh virus HPV, gejala awal ditandai dengan

Tabel 3. 1 Deskripsi Atribut *Dataset*(lanjutan)

Atribut	Deskripsi
	kemunculan kutil di berbagai bagian tubuh, termasuk alat kelamin.
STDs: Number of diagnosis	Merupakan kode nomor diagnosis bagi tiap pasien yang melakukan pemeriksaan.
STDs: Time since first diagnosis	Waktu pertama kali pasien dikenakan diagnosis mengenai penyakit menular seksual.
STDs: Time since last diagnosis	Merupakan waktu terakhir kali pasien dikenakan diagnosis penyakit menular seksual.
Dx: Cancer	Salah satu penyakit dimana sel tubuh mengalami pembelahan secara tidak normal dan merusak jaringan tubuh.
Dx: CIN	Merupakan perkembangan sel abnormal pada serviks yang dapat dilihat melalui tes Pap, HPV.
Dx: HPV	Salah satu penyakit menular seksual yang diakibatkan oleh virus HPV, ditandai dengan kemunculan kutil di berbagai bagian tubuh, termasuk alat kelamin.
Dx	Hasil diagnosis implisit yang dilakukan oleh medis.
Hinselmann	Salah satu teknik untuk melakukan pemeriksaan kanker serviks dengan mengusap serviks menggunakan kapas yang telah dimasukkan ke dalam asam asetat.
Schiller	Salah satu teknik untuk melakukan pemeriksaan kanker serviks dengan mengusap serviks menggunakan kapas yang telah dimasukkan ke dalam larutan yodium.
Citology	Prosedur medis pemeriksaan kanker dimana tahapannya dilakukan dengan cara pengambilan cairan tubuh yang akan diperiksa lebih lanjut di bawah mikroskop. *Citology* sering disebut dengan *Pap Smear*.
Biopsy	Prosedur medis pemeriksaan kanker dimana tahapannya dilakukan dengan cara pengambilan sampel jaringan yang akan diperiksa lebih lanjut di bawah mikroskop.

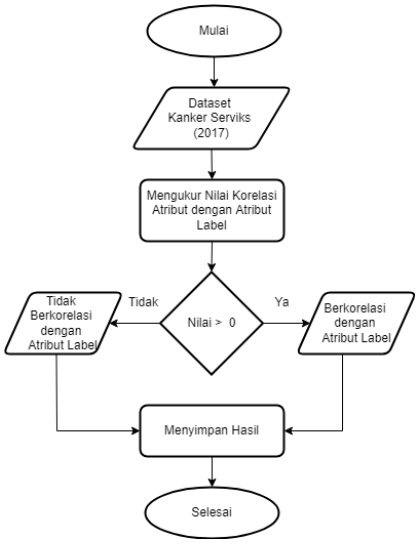
Atribut pada *dataset* dijelaskan pada Tabel sampel *record dataset* yang dipisahkan dengan koma (,). Sampel *record dataset* dapat dilihat pada Tabel 3.2.

Tabel 3. 2 Sampel Record *Dataset* Cervical Cancer (2017)

Atribut	***Dataset***						
No	1	2	3	4	…	…	858
Age	18	15	34	52	…	…	29
Number of sexual partner	4.0	1.0	1.0	5.0	…	…	2.0
First sexual intercose	15.0	14.0	?	16.0	…	…	1.0
Num of pregnancies	1.0	1.0	1.0	4.0	…	…	0.0
Smokes	0.0	0.0	0.0	1.0	…	…	0.0
Smokes (years)	0.0	0.0	0.0	37.0	…	…	0.0
Smokes (packs/years)	0.0	0.0	0.0	37.0	…	…	1.0
Hormonal Contraceptives	0.0	0.0	0.0	1.0	…	…	0.5
Hormonal Contraceptives (years)	0.0	0.0	0.0	3.0	…	…	0.0
IUD	0.0	0.0	0.0	0.0	…	…	0.0
IUD (years)	0.0	0.0	0.0	0.0	…	…	0.0
STDs	0.0	0.0	0.0	0.0	…	…	0.0
STDs (number)	0.0	0.0	0.0	0.0	…	…	0.0
STDs: condylomatosis	0.0	0.0	0.0	0.0	…	…	0.0
STDs: cervical condylomatosis	0.0	0.0	0.0	0.0	…	…	0.0
STDs: vaginal condylomatosis	0.0	0.0	0.0	0.0	…	…	0.0
STDs: syphilis	0.0	0.0	0.0	0.0	…	…	0.0
STDs: pelvic inflammatory disease	0.0	0.0	0.0	0.0	…	…	0.0
STDs: genital herpes	0.0	0.0	0.0	0.0	…	…	0.0
STDs: molluscum contagiosum	0.0	0.0	0.0	0.0	…	…	0.0
STDs: AIDS	0.0	0.0	0.0	0.0	…	…	0.0
STDs: HIV	0.0	0.0	0.0	0.0	…	…	0.0
STDs: Hepatitis B	0.0	0.0	0.0	0.0	…	…	0.0
STDs: HPV	0.0	0.0	0.0	0.0	…	…	0.0
STDs: Number of diagnosis	0.0	0.0	0.0	0.0	…	…	0
STDs: Time since first diagnosis	?	?	?	?	…	…	?
STDs: Time since last diagnosis	?	?	?	?	…	…	?
Dx: Cancer	0	0	0	1	…	…	0
Dx: CIN	0	0	0	0	…	…	0
Dx: HPV	0	0	0	0	…	…	0
Dx	0	0	0	1	…	…	0
Hinselmann	0	0	0	0	…	…	0
Schiller	0	0	0	0	…	…	0
Citology	0	0	0	0	…	…	0
Biopsy	0	0	0	0	…	…	0

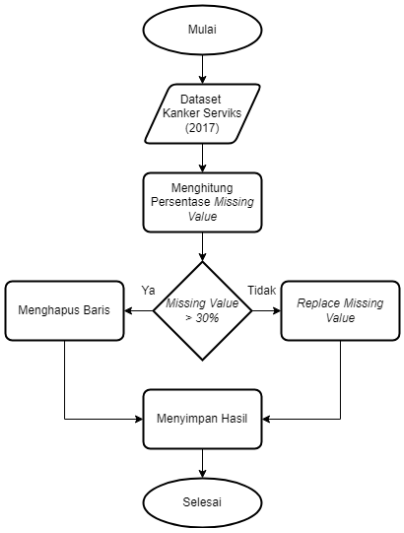
3.7 *Preprocessing Data*

Preprocessing Data pada penelitian ini dilakukan untuk menghilangkan *missing value* yang ada pada *dataset*. *Preprocessing Data* pada penelitian ini menggunakan 2 cara, yaitu Seleksi Fitur dan *Handling Missing Value*. Alur pada Seleksi fitur dapat dilihat pada Gambar 3.3.



Gambar 3. 3 Alur Seleksi Fitur

Seleksi fitur yang diambil menggunakan metode *Correlation Attribute Evaluation* untuk mengukur korelasi antara atribut label(*Biopsy*) dengan yang lain, kemudian menghapus atribut yang dianggap tidak berkorelasi dengan atribut label (*Biopsy*) [71]. *Handling Missing Value* yang dilakukan juga menggunakan 2 cara, yaitu menghapus baris dengan data *missing* bernilai lebih 30% dan *replace missing value* dengan cara imputasi menggunakan nilai rata-rata. Alur pada proses *Handling Missing Value* dapat dilihat pada Gambar 3.4.



Gambar 3. 4 Alur *Handling Missing Value*

3.7.1 *Feature Selection*

Adapun hasil dari *feature selection* adalah sebagai berikut:

1. *Dataset* yang masuk ke dalam kategori *missing value* 3,622 value dari 27,456 value yang dimasukkan ke dalam data, kurang lebih 13% - 14% dari *dataset* [69] [70]. Pada penelitian ini main target yang digunakan adalah *Biopsy* sesuai dengan penelitian [69].
2. Pada penelitian [72], 2 atribut *class* yaitu *STDs: Time Since First Diagnosis* dan *STDs: Time Since Last Diagnosis* mengandung *missing value* sebesar 91,7% serta tidak memiliki keterkaitan dengan atribut yang lain, sehingga dihapus.
3. Menggunakan metode *Correlation Atributte Eval* [71], di mana atribut dibobot dan diberi peringkat berdasarkan korelasi dan dianggap baik jika sangat berkorelasi dengan kelas. Dengan Persamaan sebagai berikut [71]:

$$Merit = \frac{\overline{kavg(corr_{fc})}}{\sqrt{k+k(k-1)\overline{avg(corr_{ff})}}} \quad (3.1)$$

Dengan metode tersebut, atribut dengan peringkat tertinggi adalah atribut nomer 34, 33, 35, 29, 31, 32, 20, 23, 30, 12, 26, 13, 17, 14, 9, 6, 1, 10, 4, 11, 5, 7, 8, 3. Sedangkan atribut nomer 22, 15, 2, 19, 21, 24, 25, 16, 28, 27, 18 memiliki peringkat rendah. Selain itu atribut dengan peringkat tinggi memiliki nilai korelasi lebih dari 0 yang berarti memiliki korelasi dengan atribut kelas utama (*Biopsy*), sedangkan atribut dengan peringkat rendah memiliki nilai korelasi kurang dari sama dengan 0 yang artinya tidak memiliki relasi dengan kelas utama (*Biopsy*). Pada Gambar 3.5 menunjukkan hasil pengolahan *dataset* metode *Correlation Atributte Eval* menggunakan software WEKA

```
Attribute Evaluator (supervised, Class (numeric): 36 Biopsy):
        Correlation Ranking Filter
Ranked attributes:
 0.7332    34 Schiller
 0.54742   33 Hinselmann
 0.32747   35 Citology
 0.1609    29 Dx:Cancer
 0.1609    31 Dx:HPV
 0.15761   32 Dx
 0.12966   20 STDs:genital herpes
 0.12413   23 STDs:HIV
 0.11317   30 Dx:CIN
 0.10674   12 STDs
 0.09745   26 STDs: Number of diagnosis
 0.09622   13 STDs (number)
 0.08698   17 STDs:vulvo-perineal condylomatosis
 0.08452   14 STDs:condylomatosis
 0.079      9 Hormonal Contraceptives (years)
 0.06148    6 Smokes (years)
 0.05596    1 Age
 0.05155   10 IUD
 0.04346    4 Num of pregnancies
 0.03225   11 IUD (years)
 0.02909    5 Smokes

 0.02466    7 Smokes (packs/year)
 0.00771    8 Hormonal Contraceptives
 0.00726    3 First sexual intercourse
 0         22 STDs:AIDS
 0         15 STDs:cervical condylomatosis
-0.00143    2 Number of sexual partners
-0.00982   19 STDs:pelvic inflammatory disease
-0.00982   21 STDs:molluscum contagiosum
-0.00982   24 STDs:Hepatitis B
-0.01389   25 STDs:HPV
-0.01967   16 STDs:vaginal condylomatosis
-0.02022   28 STDs: Time since last diagnosis
-0.02981   27 STDs: Time since first diagnosis
-0.04213   18 STDs:syphilis
```

Gambar 3. 5 Hasil *Feature Selection* Menggunakan WEKA

Total atribut yang semula berjumlah 36 akan dipakai sesuai metode ini adalah 24 atribut ditambah 1 atribut sebagai kelas utama yaitu *Biopsy*.

3.7.2 *Handling Missing Value*

Handling Missing Value akan dilakukan pada masing-masing atribut dengan cara menghapus baris dengan nilai *missing value* lebih dari 30%, mengganti *missing value* menggunakan nilai *mean* pada tiap atribut. Persamaan untuk proses penghitungan nilai *mean* sebagai berikut:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.2}$$

Proses *handling missing value* dapat dilakukan seperti tahapan dibawah ini.

1. STDs: genital herpes, STDs:HIV, Dx:CIN, STDs, STDs:number of diagnosis, STDs (number), STDs:vulvo-perineal condylomatosis, STDs:condylomatosis, nilai pada atribut dengan isi '?' diganti dengan nilai 0 karena berarti belum melakukan pemeriksaan ataupun memang bernilai 0. Tidak dapat menggunakan rata-rata, minimal-maksimal, dan modus, karena setiap orang memiliki diagnosis tersebut yang berbeda dan tingkat kesehatan yang berbeda juga.
2. Pada baris ke 73, 274, dan 763 *missing value* bernilai 45,8%, baris 76, 92, 97, 100, 104, 113, 143, 144, 145, 151, 178, 180, 195, 227, 230, 245, 258, 269, 273, 283, 286, 301, 306, 309, 318, 321, 323, 330, 332, 340, 343, 345, 348, 353, 355, 383, 393 , 407, 410, 411, 421, 438, 445, 447, 454, 455, 478, 491, 507, 514, 529, 530, 544, 549, 553, 578, 588, 605, 609, 621, 635, 737, 743, 753, 764, 769, 771, 773, 774, 780, 785, 787, 794, 796, 809, 816, dan 818 *missing value* bernilai 37,5%, baris 119, 281, dan 554 *missing value* bernilai 50%, baris 166, 222, 236, 265, 400 , 419,459, 462, 497, 546, 698, 702, 703, dan 707 *missing value* bernilai 41,7%. Hal tersebut membuat baris yang

disebutkan dihapus untuk mengurangi kesalahan pada saat eksekusi. Sehingga jumlah *dataset* yang baru adalah 756 *record*.

3. Hormonal Contraceptives (years), nilai pada atribut dengan isi "?" akan diganti dengan nilai *mean* dari total nilai atribut tersebut. Hal tersebut diambil karena dapat dilihat wanita menggunakan kontrasepsi hormon pada rata-rata usia berapa.

$$\bar{x} = \frac{1.935,648}{858} = 2,256$$

4. Smokes (years), nilai pada atribut dengan isi "?" akan diganti dengan nilai *mean* dari total nilai atribut tersebut. Hal tersebut diambil karena dapat dilihat wanita merokok pada rata-rata usia berapa.

$$\bar{x} = \frac{1.046,76}{858} = 1,22$$

5. IUD, nilai pada atribut dengan isi "?" diganti dengan nilai 0 karena berarti belum melakukan pemeriksaan ataupun memang bernilai 0. Tidak dapat menggunakan rata-rata, minimal-maksimal, dan modus, karena setiap orang belum tentu menggunakan IUD.
6. Smokes, nilai pada atribut dengan isi "?" diganti dengan nilai 0 karena berarti belum melakukan pemeriksaan ataupun memang bernilai 0. Tidak dapat menggunakan rata-rata, minimal-maksimal, dan modus, karena setiap orang belum tentu merokok.
7. Smokes (packs/years), nilai pada atribut dengan isi "?" diganti dengan nilai *mean* dari total nilai atribut tersebut. Hal tersebut diambil karena dapat dilihat rata-rata jumlah rokok yang dipakai wanita pertahun.

$$\bar{x} = \frac{388,674}{858} = 0,453$$

8. Hormonal Contraceptives, nilai pada atribut dengan isi "?" diganti dengan nilai 0 karena berarti belum melakukan pemeriksaan ataupun memang bernilai 0. Tidak dapat menggunakan rata-rata, minimal-maksimal, dan modus, karena setiap orang belum tentu menggunakan Hormonal Contraceptives.

9. First Sexual Intercouse, nilai pada atribut dengan isi "?" diganti dengan nilai *mean* dari total nilai atribut tersebut. Hal tersebut diambil karena dapat dilihat rata-rata wanita melakukan hubungan seksual pertama kali pada umur berapa.

$$\bar{x} = \frac{14.581,71}{858} = 16,995$$

10. Num of Pregnancies, nilai pada atribut dengan isi "?" diganti dengan nilai *mean* dari total nilai atribut tersebut. Hal tersebut diambil karena dapat dilihat rata-rata jumlah kehamilan wanita.

$$\bar{x} = \frac{1.952,808}{858} = 2,276$$

11. IUD (years), nilai pada atribut dengan isi "?" akan diganti dengan nilai *mean* dari total nilai atribut tersebut. Hal tersebut diambil karena dapat dilihat wanita memasang IUD pada rata-rata berapa tahun.

$$\bar{x} = \frac{441,87}{858} = 0,515$$

3.7.3 *Dataset* Baru

Pada Tabel 3.3 merupakan tabel yang berisi *dataset* yang telah dilakukan *preprocessing*.

Tabel 3. 3 *Dataset* Baru Hasil *Preprocessing*

Atribut	***Dataset***						
No	1	2	3	4	…	…	756
Age	18	15	34	52	…	…	29
First sexual intercose	15.0	14.0	16.995	16.0	…	…	20.0
Num of pregnancies	1.0	1.0	1.0	4.0	…	…	1.0
Smokes	0.0	0.0	0.0	1.0	…	…	0.0
Smokes (years)	0.0	0.0	0.0	37.0	…	…	0.0
Smokes (packs/years)	0.0	0.0	0.0	37.0	…	…	0.0
Hormonal Contraceptives	0.0	0.0	0.0	1.0	…	…	1.0
Hormonal Contraceptives (years)	0.0	0.0	0.0	3.0	…	…	0.5
IUD	0.0	0.0	0.0	0.0	…	…	0.0
IUD (years)	0.0	0.0	0.0	0.0	…	…	0.0
STDs	0.0	0.0	0.0	0.0	…	…	0.0
STDs (number)	0.0	0.0	0.0	0.0	…	…	0.0
STDs: condylomatosis	0.0	0.0	0.0	0.0	…	…	0.0
STDs: genital herpes	0.0	0.0	0.0	0.0	…	…	0.0
STDs: HIV	0.0	0.0	0.0	0.0	…	…	0.0

Tabel 3. 3 *Dataset* Baru Hasil *Preprocessing* (lanjutan)

Atribut	***Dataset***						
STDs: Number of diagnosis	0.0	0.0	0.0	0.0	…	…	0
Dx: Cancer	0	0	0	1	…	…	0
Dx: CIN	0	0	0	0	…	…	0
Dx: HPV	0	0	0	0	…	…	0
Dx	0	0	0	1	…	…	0
Hinselmann	0	0	0	0	…	…	0
Schiller	0	0	0	0	…	…	0
Citology	0	0	0	0	…	…	0
Biopsy	0	0	0	0	…	…	0

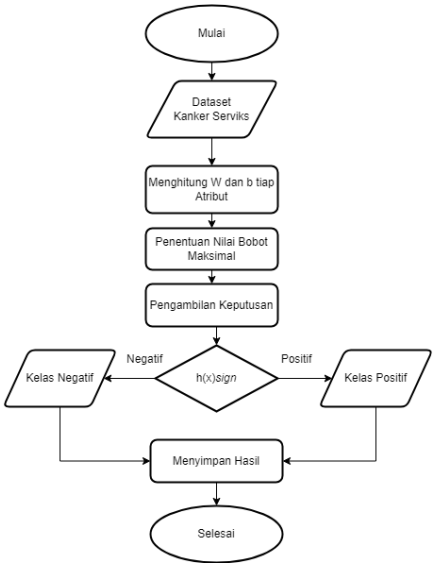
3.8 10-Fold Cross Validation

Penelitian ini menggunakan pembagian data latih dan uji dengan menggunakan metode *Cross Validation* dengan nilai k=10.

3.9 Alur Algoritma

3.9.1 *Support Vector Machine*

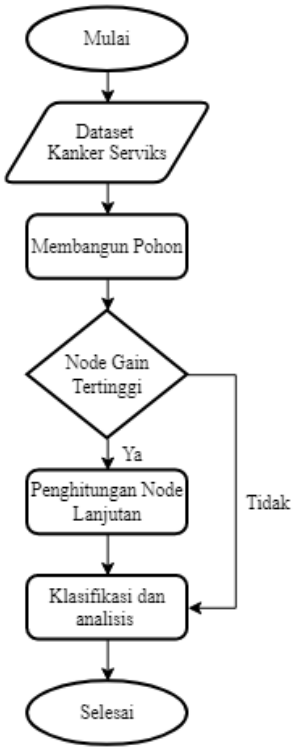
SVM adalah salah satu algoritma metode klasifikasi yang memiliki beberapa tahapan dalam proses eksekusi. Alur algoritma *Support Vector Machine* dapat dilihat pada Gambar 3.6



Gambar 3. 6 Diagram Alur Algoritma SVM

3.9.2 *Random Forest*

Random Forest merupakan salah satu algoritma gabungan dari beberapa *decision tree* yang memiliki beberapa tahapan dalam proses eksekusinya. Alur algoritma *Random Forest* dapat dilihat pada Gambar 3.7.



Gambar 3. 7 Diagram Alur Algoritma *Random Forest*

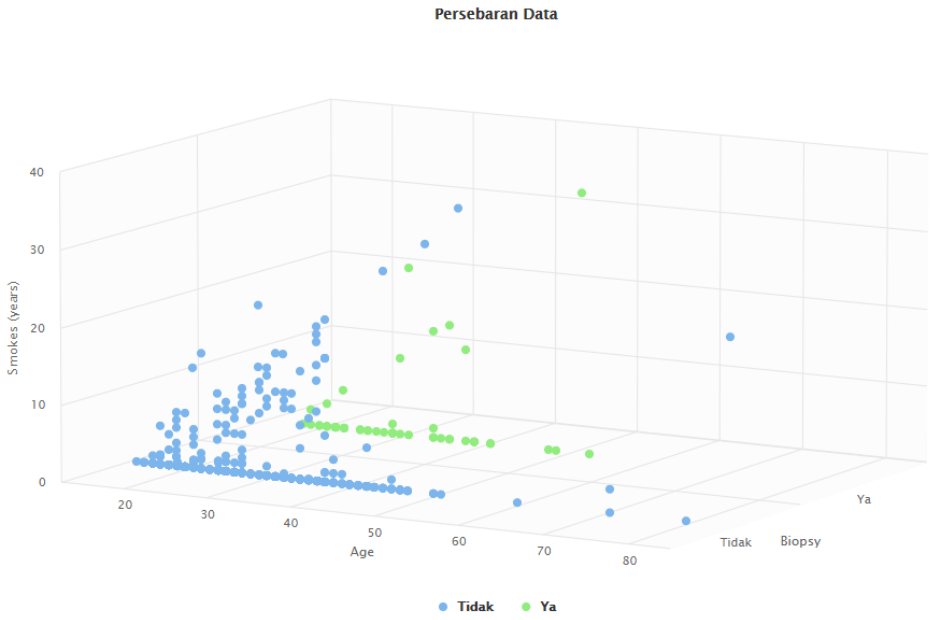
3.10 Pelatihan Data

Pelatihan data dilakukan dengan melatih *dataset* yang sudah dibagi atau ditentukan agar dapat menjalankan fungsi untuk memprediksi hasil dari klasifikasi menggunakan algoritma SVM dan RF.

3.11 Analisis Data

Pada proposal penelitian ini dilakukan proses klasifikasi menggunakan dua algoritma, yaitu *Random Forest* dan *Support Vector Machine*. Kemudian hasil dari klasifikasi akan dibandingkan dari segi akurasi, proses komputasi, dan tingkat *error* untuk menentukan algoritma mana yang memiliki performa terbaik. Hasil *dataset* yang telah diolah sesuai dengan Tabel 3.3 dikenai eksperimen awal menggunakan

software rapidminer dengan melakukan klasifikasi terhadap data tersebut menggunakan algoritma SVM dan RF. Pembagian data latih dan uji pada penelitian ini dilakukan menggunakan cara *k-cross validation* dengan nilai k=10. Persebaran data kanker serviks dapat dilihat pada Gambar 3.8.



Gambar 3. 8 Persebaran Data Kanker Serviks

Dapat dilihat pada Gambar 3.8 bahwa data kanker serviks berbentuk *non-*linear karena data tersebar dan tidak terbentuk menjadi satu garis linear. Algoritma SVM yang dipakai menggunakan kernel RBF gaussian sesuai dengan penelitian [73][74], bahwa kernel tersebut adalah kernel terbaik untuk melakukan klasifikasi terhadap data *non-linear*. Pada penelitian ini, nilai sigma gamma 1,2, dan 3 adalah 1.0. Eksperimen awal dilakukan dengan menggunakan *tools*, yaitu software rapidminner. Hasil dari pengolahan data menggunakan algoritma SVM dapat dilihat pada Tabel 3.5.

Tabel 3. 4 Hasil Performa Algoritma SVM

	***True* Tidak**	***True* Ya**	***Class Precission***
Pred. Tidak	702	54	92.86%
Pred. Ya	0	0	0.00%
Class Recall	100.00%	0.00%	
Accuracy	92.86%		

Tabel 3. 5 Hasil Klasifikasi dengan Algoritma SVM

Kelas	Prediksi Benar	Nilai Sebenarnya
Tidak	702	712
Ya	0	44

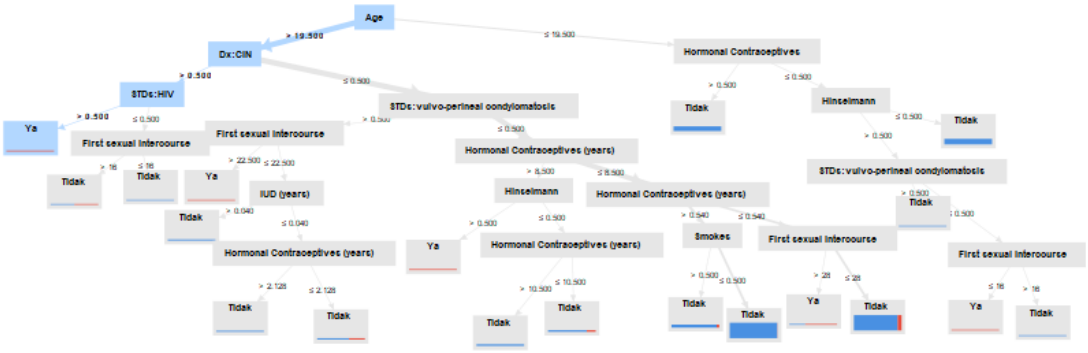
Pada Tabel 3.6 dapat dilihat bahwa algoritma SVM pada kelas tidak, tidak dapat memprediksi semua data yaitu 702 dari 712 data yang artinya 10 data tidak terbaca. Kemudian untuk kelas ya, tidak ada data yang diprediksi dengan benar. Hasil dari klasifikasi dengan SVM secara keseluruhan dapat dilihat pada link : https://bit.ly/HasilKlasifikasiSVM.

Penggunaan algoritma RF dapat dilihat pada Tabel 3.7. Algoritma RF yang dipakai pada pengujian ini menggunakan *number of trees* = 10 dan *maximal depth* terbaik yang telah diuji dari rentang 1 sampai dengan 10 adalah 7 dengan penetapan node dari nilai *information gain*. Percobaan *maximal depth* dengan 10 kali percobaan dapat dilihat pada Tabel 3.7.

Tabel 3. 6 Pencarian *Maximal Depth* Terbaik

percobaan ke	***max_depth***	**Akurasi**
1	1	92,86%
2	2	93,66%
3	3	93,66%
4	4	94,05%
5	5	93,79%
6	6	94,05%
7	7	94,45%
8	8	94,32%
9	9	93,80%
10	10	93,53%

Salah satu *tree* yang dihasilkan pada percobaan dengan algoritma *Random Forest* dapat dilihat pada Gambar 3.9.



Gambar 3. 9 *Tree* pada Pengujian dengan *Random Forest*

Tabel 3.8 menunjukkan hasil dari pengujian menggunakan algoritma RF.

Tabel 3. 7 Hasil Performa Algoritma RF

	***True* Tidak**	***True* Ya**	***Class Precission***
Pred. Tidak	686	26	96.35%
Pred. Ya	16	28	63.64%
Class Recall	97.76%	51.85%	
Accuracy	94.45%		

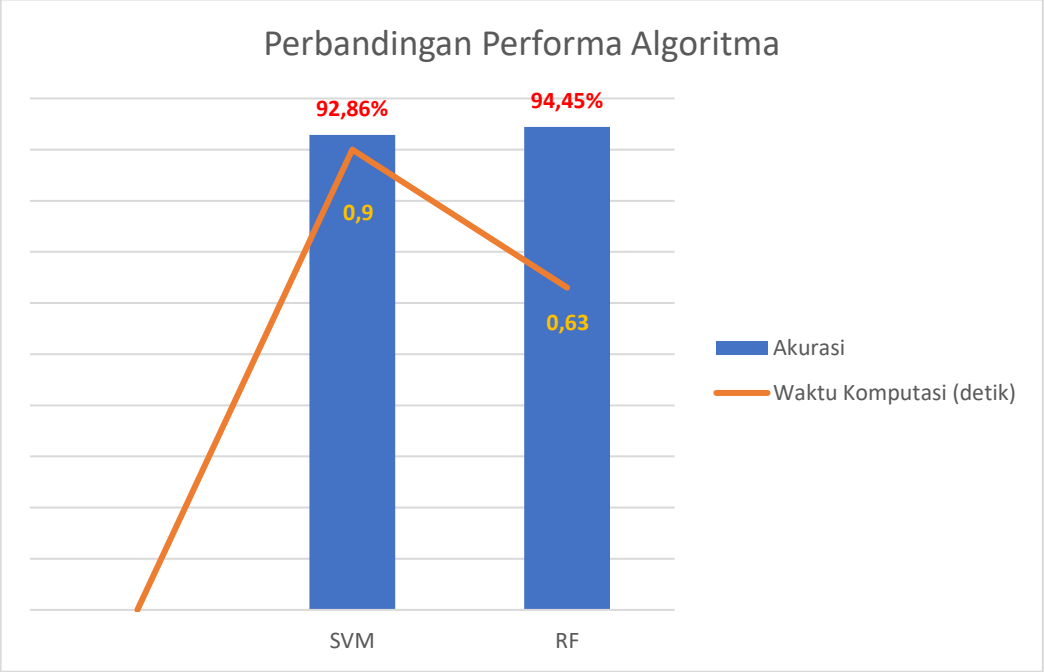
Tabel 3.9 menunjukkan hasil klasifikasi menggunakan algoritma RF.

Tabel 3. 8 Hasil Klasifikasi dengan Algoritma RF

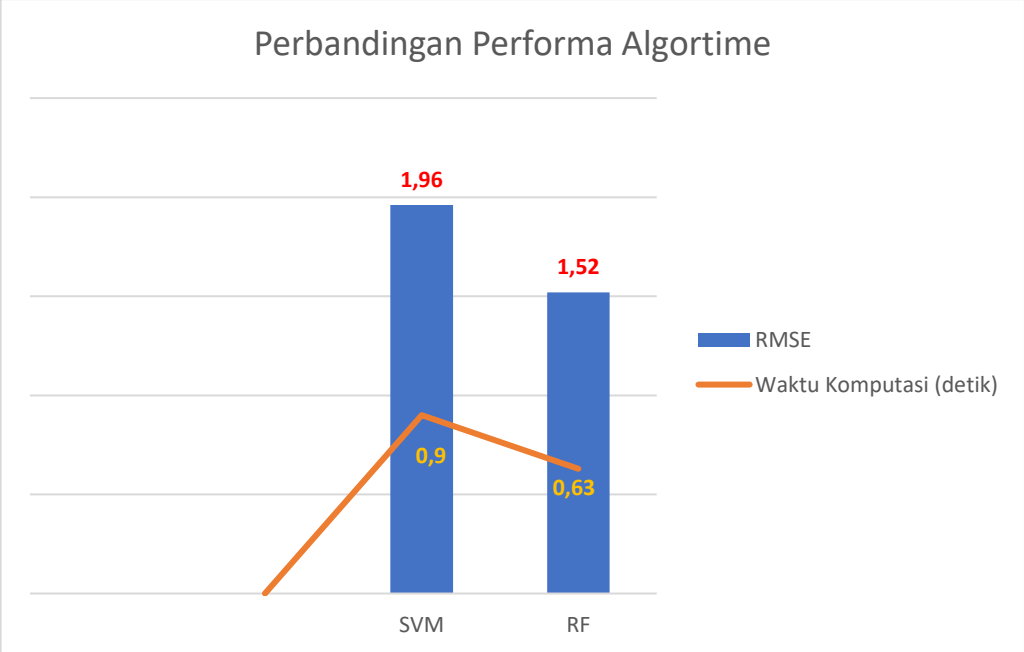
Kelas	Prediksi Benar	Nilai Sebenarnya
Tidak	686	712
Ya	28	44

Pada Tabel 3.9 dapat dilihat bahwa RF dapat memprediksi seluruh data yang ada yaitu 756 data. Hasil klasifikasi dengan RF secara keseluruhan dapat dilihat pada link: https://bit.ly/HasilKlasifikasiRandomForest.

Visualisasi dari hasil perbandingan dapat dilihat pada Gambar 3.10 yang membandingkan algoritma berdasar akurasi dan waktu komputasi serta Gambar 3.11 berdasarkan RMSE dan waktu komputasi.



Gambar 3. 10 Perbandingan Algoritma Berdasarkan Akurasi dan Waktu Komputasi



Gambar 3. 11 Perbandingan Algoritma Berdasarkan RMSE dan Waktu Komputasi

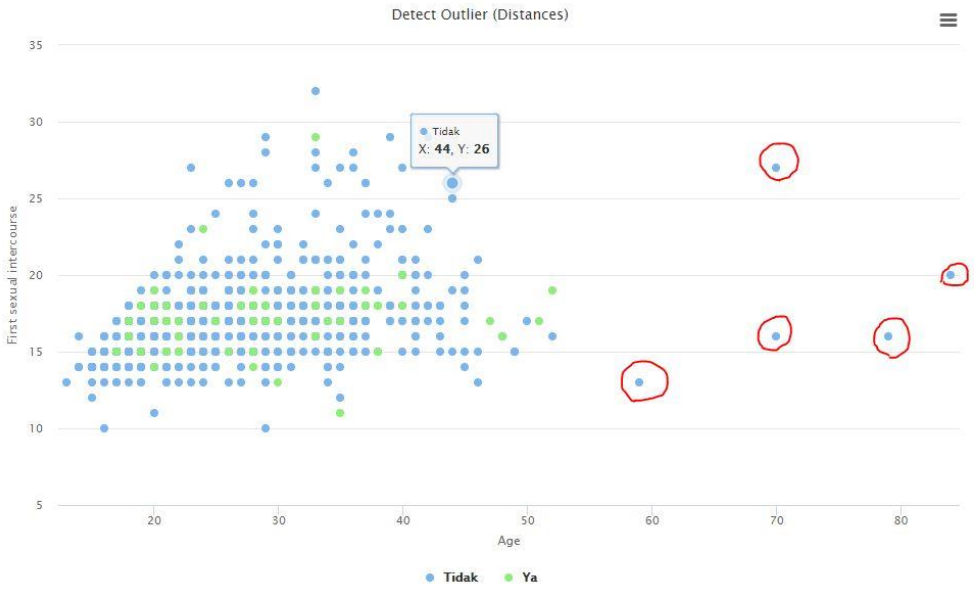
Berdasarkan Gambar 3.10 Algoritma RF lebih unggul, dimana waktu komputasinya lebih rendah sehingga menghasilkan akurasi yang lebih tinggi. Gambar 3.11 menunjukkan bahwa saat nilai RMSE lebih rendah yaitu pada algoritma RF, maka waktu komputasi yang dibutuhkan akan rendah juga karena *error* yang dihasilkan juga kecil. Perbandingan performa algoritma secara menyeluruh dapat dilihat pada Tabel 3.10 dengan nilai MSE, dan RMSE dapat dihitung menggunakan persamaan 2.23. dan 2.24. Kemudian untuk waktu komputasi dilakukan perhitungan menggunakan stopwatch pada smarthphone karena pada aplikasi rapidminer tidak dapat menghitung waktu komputasi dengan nilai desimal.

Tabel 3. 9 Hasil Perbandingan Algoritma SVM dan RF

Parameter	**SVM**	**RF**
Recall Kelas Tidak	100%	97.76%
Recall Kelas Ya	0%	51,85%
Precission Prediksi Tidak	92,86%	96.35%
Precission Prediksi Ya	0%	63.64%
`MSE	1.458	882
RMSE	1,96	1,52
Akurasi	92,86%	94,45%
Waktu Komputasi	0,90 detik	0,63 detik

Tabel 3.10 menunjukkan bahwa algoritma RF memiliki akurasi yang lebih baik daripada SVM. Kemudian untuk tingkat kesalahan dalam eksekusi dapat dilihat pada nilai MSE dan RMSE, apabila parameter tersebut memiliki nilai yang semakin kecil, maka akan semakin besar akurasi yang dihasilkan. Terbukti dengan akurasi algoritma RF lebih tinggi daripada SVM karena memiliki nilai MSE dan RMSE lebih kecil daripada SVM. Pada saat proses klasifikasi SVM tidak dapat memprediksi seluruh data yang ada sedangkan RF dapat memprediksi semua data. SVM juga memiliki tingkat kesalahan yang tinggi pada kelas ya yang semua datanya diprediksi salah. Waktu komputasi dari kedua algoritma tersebut juga terlihat lebih unggul pada algoritma RF dibandingkan dengan SVM. Dapat disimpulkan bahwa algoritma yang memiliki performa lebih unggul pada percobaan awal menggunakan *rapidminer* adalah dibanding dengan algoritma SVM. Pada RF

proses pengambilan keputusan dibantu dengan adanya penggunaan *tree* dan *max depth* terbaik sehingga membuat akurasi semakin tinggi. Sedangkan pada SVM, data bersifat *non-linear* dalam pengolahannya lebih kompleks karena data tersebar dan memungkinkan adanya *outlier* sehingga menghasilkan akurasi yang lebih rendah. *Outlier* adalah data yang berbeda dari data lainnya dapat dilihat pada persebaran data jika data tersebut berada diluar sebaran data[75]. Terbukti saat dilakukan pengecekan dengan rapidminer *dataset* tersebut memiliki *outlier*. *Outlier* pada data dapat dilihat pada Gambar 3.12.



Gambar 3. 12 Pengecekan *Outlier*