BAB II
TINJAUAN PUSTAKA

2.1 Penelitian terdahulu

Salah satu faktor penyebab kematian wanita di dunia adalah kanker serviks. Menurut penelitian [29], data hasil rekam medis pasien penyakit kanker serviks tidak disertai dengan proses pengolahan yang baik untuk menghasilkan suatu informasi. Proses pengklasifikasian data *pap smear* terhadap studi kasus kanker menggunakan *data mining* memiliki potensi yang tinggi. Data yang digunakan adalah data sesudah pasien didiagnosis terjangkit kanker sehingga tidak dapat mendeteksi sebelum pasien terjangkit. Kemudian pada penelitian ini proses *preprocessing* data belum cukup baik dan *dataset* yang digunakan masih dalam jumlah sedikit. Pada penelitian [30][31][32], kanker serviks adalah salah satu penyakit mematikan yang dapat diatasi dengan cara mengklasifikasikan faktor risiko untuk dapat mendiagnosis secara dini. Penelitian tersebut menyatakan bahwa terdapat beberapa faktor yang menyebabkan seseorang berisiko mengalami kanker serviks yaitu, usia penderita, hubungan seksual pertama kali yang dilakukan, jumlah kehamilan, kebiasaan merokok, kontrasepsi yang bersifat hormonal, PMS herpes genital, dan lainnya.

Ada banyak algoritma *machine learning* yang dapat menangani proses pengklasifikasian. Menurut penelitian [33][23][34], algoritma SVM merupakan algoritma dengan performa yang baik. Saat dibandingkan dengan algoritma NN, SVM memiliki akurasi yang lebih baik yaitu 95,16% dibanding dengan NN sebesar 93,37%. Kemudian SVM memiliki hasil paling sesuai yang diharapkan serta jika dibandingkan dengan algoritma regresi logistic biner, SVM memiliki kesalahan proses klasifikasi lebih kecil.

Sedangkan menurut penelitian [35][36][37], *random forest* adalah algoritma dengan performa yang baik. Saat dilakukan perbandingan dengan algoritma *Regresi*

Linear, Random Forest memiliki akurasi 97,7% sedangkan *Regresi Linear* sebesar 94%. Kemudian jika dibandingkan ANN, *Random Forest* memiliki akurasi sebesar 90,62%, sedangkan ANN sebesar 82,29%. Lalu jika dibandingkan dengan algoritma C.45 dan *Gradient Boosting*, *Random Forest* memiliki akurasi tertinggi sebesar 90%, diikuti C.45 sebesar 86,67%, dan terakhir *Gradient Boosting* sebesar 75%.

Tabel 2. 1 Penelitian Terkait Kanker dan Algoritma Klasifikasi

No	Judul	Peneliti	Masalah	Metode dan hasil (akurasi)
1	Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, I Bayes, dan k-NN	Tutus Praningki dan I Budi (2018) [29].	Pada tahun 2013, di Indonesia penderita kanker serviks lebih banyak jumlahnya dibandingkan kanker payudara. Khususnya di provinsi Jawa Timur dengan jumlah kasus 21.313. Hal tersebut membuat kanker serviks perlu diklasifikasikan.	Metode klasifikasi dengan algoritma CART, KNN, dan Naïve Bayes. Hasil akurasi, Naïve Bayes tertinggi sebesar 94,44%, diikuti CART sebesar 88,89%, dan terakhir KNN sebesar 85,04%. Penelitian ini berfokus pada dataset hasil pap smear hasil wawancara di RSUD Kediri, tidak memprediksi awal sebelum kanker menjangkit.
2	Cervical cancer classification using convolutional neural networks and extreme learning machines	Ahmed Ghoneim dkk 9 (2019) [30].	Salah satu penyebab utama kematian wanita didunia adalah karena terjangkit kanker serviks. Masalah ini dapat diatasi jika kanker serviks dapat didiagnosis dan diobati semenjak stadium dini	Diagnosis kanker serviks dapat dilakukan dengan menggunakan metode klasifikasi salah satunya dengan CNN dan ELM terhadap dataset berbentuk gambar.
3.	Classification Of Cervical Cancer Using Convolutional Neural Network (Alexnet)	Habibullah Akbar dan Sandfreni [31].	Menurut Kementrian Kesehatan Indonesia, pada tahun 2018 terdapat sebesar 23,4 persen kasus kanker serviks di Indonesia dari 100.000 penduduk dengan total kematian sebanyak 18.279 jiwa pertahunnya.	Proses diagnosis penyakit kanker serviks dapat dilakukan dengan metode pengklasifikasian menggunakan algoritma CNN terhadap dataset Intel dan

Tabel 2. 1 Penelitian Terkait Kanker dan Algoritma Klasifikasi

No	Judul	Peneliti	Masalah	Metode dan hasil (akurasi)
				MobileODT Cervical Cancer Screening.
4	Cervical Cancer Diagnosis Using Random Forest Classifier With SMOTE and Feature Reduction Techniques	Sherif F. Abdoh dkk (2018) [32].	Kanker serviks adalah penyakit ganas yang paling umum terjadi pada wanita di seluruh dunia. Di sebagian besar kasus, gejala kanker serviks tidak terlihat pada tahap awal. Ada banyak faktor yang menyebabkan risiko terkena kanker serviks seperti human papilloma virus, penyakit menular seksual, dan merokok.	Dengan metode klasifikasi menggunakan RF, dapat diketahui atribut paling penting yang mempengaruhi terjangkitnya kanker serviks yaitu atribut nomer 10, 8, 12, dan 3.
5	Perbandingan Metode Data Mining SVM Dan NN Untuk Klasifikasi Penyakit Ginjal Kronis.	Hilda Amalia (2018) [33].	Penyakit kronis ginjal semakin hari semakin memprihatinkan. Menurut data WHO, Indonesia diperkirakan mengalami peningkatan sebesar 46% dari tahun 1955-2025. Penyakit mematikan lainnya dapat disebabkan oleh penyakit ginjal kronik, diantaranya lupus, darah tinggi, dan jantung.	Metode klasifikasi dengan algoritma SVM dan NN. SVM memiliki akurasi lebih baik dari NN yaitu sebesar 95,16%, sedangkan NN sebesar 93,37%. Dataset yang digunakan diambil dari website UCI dengan bentuk supervised.
6	Perbandingan Algoritma SVM Dan KNN dalam Mengklasifikasi Kelulusan Mahasiswapada Suatu Mata Kuliah.	Shedriko (2021) [23].	Universitas yang mahal belum tentu berkualitas dan universitas yang memiliki kualitas baik belum tentu mahal. Kualitas pendidik menjadi salah satu faktor penting penentu kualitas suatu universitas.	Metode klasifikasi dengan algoritma KNN dan SVM. Hasil yang didapat mengatakan bahwa algoritma yang lebih unggu adalah KNN, tetapi algoritma SVM memiliki hasil klasifikasi yang lebih sesuai dengan yang diharapkan.
7	Perbandingan Kinerja Klasifikasi Support Vector	I.T. Utami (2018) [34].	Mahasiswa lebih banyak menyelesaikan	Metode klasifikasi dengan algoritma

Tabel 2. 1 Penelitian Terkait Kanker dan Algoritma Klasifikasi (lanjutan)

No	Judul	Peneliti	Masalah	Metode dan hasil (akurasi)
	Machine (Svm) Dan Regresi Logistik Biner Dalam Mengklasifikasikan Ketepatan Waktu Kelulusan Mahasiswa Fmipa Untad		studi sarjana selama lebih dari 4 tahun yang dapat menyebabkan penurunan kualitas penilaian pada perguruan tinggi.	Regresi Logistik Biner dan SVM.Algoritma SVM dapat membuktikan bahwa algoritma tersebut memiliki kesalahan kecil dalam melakukan klasifikasi dibanding Regresi Logistik Biner.
8	Perbandingan Algoritma Regresi Linier dan Regresi Random Forest Dalam Memprediksi Kasus Positif Covid-19	Syakirah Fachid dan Agung Triyaudi (2022) [35].	Pandemi COVID-19 berdampak besar pada kehidupan masyarakat di seluruh dunia apalagi dengan kebijakan *lockdown.* Semua bidang, seperti ekonomi, pariwisata, pendidikan, dan masih banyak lagi mengalami banyak kerugian karena dampak dari pandemic COVID-19.	Metode klasifikasi denga algoritma random forest dan regresi linear. Akurasi yang dihasilkan random forest lebih baik dari regresi linear yaitu sebesar 97,7%, sedangkan regresi linear sebesar 94%.
9	Perbandingan Akurasi Algoritma Naïve Bayes Dan Algoritma Artificial Neural Network untuk Klasifikasi Penyakit Diabetes	Muhammad Kaddafi Nasution (2021) [36].	Indonesia menduduki peringkat ke 7 dari 10 besar negara dengan jumlah pasien diabetes tertinggi di dunia. Hal tersebut didasarkan oleh data *International Diabetes Federation* (IDF). Identifikasi yang dapat dilakukan terhadap penyakit tersebut bisa menggunakan proses pengklasifikasian penyakit diabetes	Metode klasifikasi dengan algoritma random forest dan ANN. Akurasi random forest lebih baik dari ANN yaitu sebesar 90,62%, dibandingkan dengan ANN yang sebesar 82,29%.
10	Komparasi Kinerja Algoritma C4.5, Random Forest, dan Gradient Boosting untuk Klasifikasi Komoditas	Edi Ismanto dan Melly Novalia (2021) [37].	Permasalahan sektor perekonomian di provinsi Riau yaitu ketahanan pangan, kemiskinan, dan pembangunan yang tertinggal. Dari hal tersebut diperlukan klasifikasi untuk melihat pola untuk	Metode klasifikasi dengan algoritma C.45, Random Forest, dan *Gradient Boosting*. Hasil akurasi, tertinggi sebesar 90%, diikuti C.45 sebesar 86,67%, dan terakhir

Tabel 2. 1 Penelitian Terkait Kanker dan Algoritma Klasifikasi (lanjutan)

No	Judul	Peneliti	Masalah	Metode dan hasil (akurasi)
	mempermudah mendapat informasi mengenai komoditi unggulan.			*Gradient Boosting* sebesar 75%.

Dilihat dari penelitian [29][30][31][32], kanker serviks merupakan salah satu penyakit berbahaya urutan kedua penyebab kematian wanita di dunia. Oleh karena itu, dibutuhkan langkah untuk dapat mengatasi secara dini agar penyakit tersebut dapat dihindari. Perbedaan penelitian terkait tersebut dengan penelitian yang dilakukan adalah, pada penelitian [29] dataset berjumlah kecil dan belum diolah secara baik, penelitian [30][31] dataset yang digunakan berbeda dengan yang akan digunakan penulis. Penulis menggunakan dataset berbentuk supervised karena proses yang akan dilakukan adalah klasifikasi. Menurut penelitian [33][23][34], algoritma SVM memiliki performa baik dilihat dari akurasi tinggi ketika dibandingkan dengan algoritma yang lain. Selain itu SVM memiliki kesalahan eksekusi kecil dengan keluaran yang dihasilkan sesuai harapan. Kemudian menurut penelitian [35][36][37], algoritma RF memiliki performa terbaik dilihat dari akurasi yang tinggi jika dibandingkan dengan algoritma yang lain. Maka dikarenakan hal tersebut, penulis memilih algoritma SVM dan RF untuk dibandingkan agar mendapatkan algoritma dengan performa paling terbaik dari keduanya.

2.2 Dasar Teori

2.2.1 Kanker Serviks

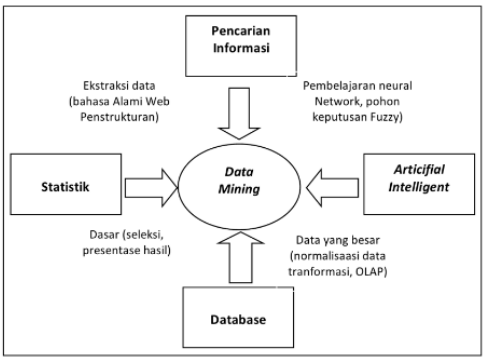
Serviks merupakan bagian yang terletak pada bawah rahim yang memiliki fungsi sebagai jalan kelahiran dan memisahkan antara vagina dan rahim [38]. Penyakit yang dapat menjangkit pada bagian serviks salah satunya adalah kanker [38]. Kanker serviks adalah salah satu dari sekian banyak jenis kanker yang diakibatkan oleh virus bernama *Human Papilloma Virus* (HPV) onkogenik. Virus tersebut menjangkit bagian leher rahim pada wanita subur kisaran 15-49 tahun [39]. HPV sub tipe onkogenik merupakan penyebab

utama kanker serviks, selain itu ada beberapa faktor yang menyebabkan seseorang memiliki risiko kanker serviks diantaranya aktivitas seksual diusia muda, banyaknya partner aktivitas seksual, merokok, jumlah anak, keadaan ekonomi menengah kebawah, pemakaian kontrasepsi seperti pil KB, penyakit menular seksual, adanya gangguan pada imunitas, dan faktor genetik [40]. Kanker serviks dapat dilihat dan didiagnosa dengan melihat beberapa indikator atau pemeriksaan, diantaranya:

- *Cytology*, merupakan upaya sekunder yang berfungsi untuk melakukan pendeteksian secara dini kanker serviks dengan metode sitologi serviks [41].
- *Biopsy*, merupakan teknik pemeriksaan kanker dengan pengambilan cairan pada jaringan yang terjangkit kanker dengan FNA yang kemudian hasilnya akan didiagnosis [42].
- *Hinselman (IVA-Test)*, merupakan teknik pemeriksaan dengan melakukan inspeksi visual asam asetat yang kemudian hasilnya akan didiagnosis [40].
- *Schiller*, merupakan teknik pemeriksaan dengan melakukan inspeksi cairan iodium ke dalam serviks yang kemudian hasilnya akan didiagnosis [28].

2.2.2 Data Mining

Data mining adalah salah satu cabang ilmu yang merupakan aktivitas untuk pengumpulan data guna mendapatkan pengetahuan dan informasi dalam data berukuran besar, dimana hasilnya akan digunakan sebagai alternatif dalam pengambilan suatu keputusan [43]. *Data mining* berasal dari berbagai disiplin ilmu dan yang paling berkontribusi adalah ilmu statistika dan *artificial intelligent* [44]. Salah satu teknik pada *data mining* yaitu dengan melakukan pengamatan data yang didapat dan kemudian dibangun sebuah model untuk mengenali data tersebut sehingga pola *universal* pada data dapat diketahui [45]. Pada Gambar 2.1 terdapat 4 bidang ilmu dalam *data mining*.



Gambar 2. 1 Bidang Data Mining

(Sumber : Konsep Data Mining Vs Sistem Pendukung Keputusan[46])

Data mining dapat dikelompokkan berdasarkan tugas yang dilakukan, yaitu [47]:

1. Deskripsi

   Deskripsi merupakan salah satu tahapan yang digunakan untuk mencari teknik untuk menggambarkan pola dan kecenderungannya terhadap suatu data.

2. Estimasi

   Estimasi hampir sama dengan klasifikasi tetapi memiliki perbedaan pada jenis variabel yang lebih mengarah pada numerik. Selain itu model akan dibangun menggunakan *record* lengkap yang akan dijadikan sebagai tolak ukur prediksi.

3. Prediksi

   Prediksi juga dapat dikatakan hampir sama dengan klasifikasi dan estimasi, tetapi berbeda pada hasil yang menunjukkan akan terjadi pada masa yang akan datang.

4. Klasifikasi

   Klasifikasi merupakan tugas yang dilakukan berdasar target variabel kategori sebagai kelas pembeda yang akan menjadi tolak ukur.

5. Pengklusteran

Pengklusteran adalah tahapan dalam mengamati atau mengobservasi suatu objek atau data agar terbentuk suatu kelas yang memiliki kesamaan antara satu dengan yang lainnya.

6. Asosiasi

Asosiasi merupakan tahapan yang digunakan untuk menemukan atribut pada jangka satu waktu dan sering disebut dengan analisis keranjang belanja *data mining*.

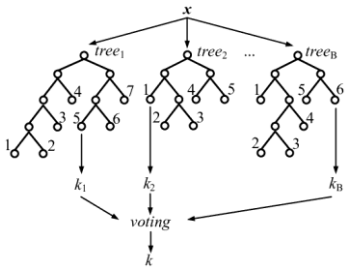
2.2.3 *Data Preprocessing*

Preprocessing adalah salah satu proses awal metode *data mining* untuk menghasilkan analisis yang lebih bagus dan akurat. Selain itu *preprocessing* bermanfaat untuk mempersingkat waktu eksekusi dan dapat membuat data berukuran lebih kecil tanpa melakukan pengurangan isi yang ada di dalamnya [48]. *Preprocessing data* dapat dilakukan salah satunya dengan menggunakan metode *Feature Selection,* Seleksi fitur terdiri dari 2 tahap [49], tahap pertama dilakukan untuk melakukan pelatihan dan pengujian. Sedangkan tahap kedua adalah penilaian kemampuan dengan metode pemilihan fitur (atribut) untuk menerapkan urutan seberapa penting fitur tersebut untuk proses kalsifikasi. Tahapan umum lainnya yang sering dilakukan adalah imputasi, yaitu dengan mengisi data yang kosong dengan nilai layak [50]. Teknik pada imputasi juga beragam, salah satunya adalah menggunakan *mean*, yaitu melakukan proses penggantian data yang kosong dengan nilai rata-rata dari keseluruhan data yang ada dan juga terdapat metode *median*, yaitu dengan menghitung nilai tengah dari keseluruhan data yang ada [50].

2.2.4 Algoritma Random Forest

Random Forest merupakan salah satu dari banyaknya algoritma klasifikasi dengan jenis *Supervised Learning* yang menggunakan metode

percabangan pohon atau *Decision Tree* [51]. Tiap cabang pada Random Forest memiliki pernyataan masing-masing untuk memecahkan masalah demi mencapai sebuah keputusan final [52]. Algoritma Random Forest melakukan pembentukan pohon dengan cara pelatihan sampel data dengan pemilihan variabel acak yang kemudian penentuan klasifikasi diambil berdasar nilai dari vote masing-masing pohon yang terbentuk [53]. Bentuk arsitektur umum dari algoritma Random Forest dapat dilihat pada Gambar 2.2.



Gambar 2. 2 Arsitektur Umum Random Forest

(Sumber : Mining data with random forests: A survey and results of new tests [54])

Pohon keputusan pada algoritma *random forest* dimulai dengan melakukan perhitungan terhadap nilai *entropy* yang nantinya akan digunakan untuk melakukan perhitungan nilai *information gain*. Persamaan 2.1 dan 2.2 menunjukkan rumus *entropy* dan *information gain* [55].

$$Entropy\ (Y) = -\sum_i p(c|Y) Log_2 p(c|Y) \tag{2.1}$$

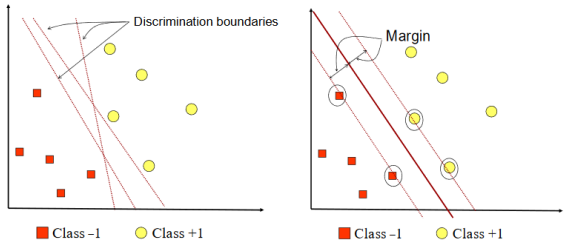
Dimana Y merupakan himpunan kasus dan p(c|Y) adalah rasio nilai Y terhadap c.

$$Information\ Gain(Y, a) = Entropy(Y) - \sum_{veValues} \frac{|Yv|}{|Ya|} Entropy(Yv) \tag{2.2}$$

Di mana nilai (a) merupakan seluruh nilai yang terjadi. Yv merupakan *subclass* dari Y, dan kelas v sesuai dengan kelas a. Ya merupakan nilai yang sinkron untuk semua kelas.

2.2.5 Algoritma *Support Vector Machine*

Support Vector Machine merupakan algoritma yang pertama kali dikenalkan oleh Vapnik pada tahun 1992 sebagai jaringan konsep dibidang pengenalan pola[56]. SVM merupakan salah satu algoritma *Machine Learning* yang memiliki kerja yang memiliki prinsip *Structure Risk Minimization* (SRM), dengan tujuan ditemukannya *hyperplane* terbaik yang dapat melakukan pemisahan dua buah kelas [57]. Kelebihan dari algoritma SVM adalah memiliki kemampuan menerapkan pemisah linier dan *nonlinear* yang diperoleh dengan fungsi kernel yang diperlukan [58]. Teori SVM diawali dengan mengelompokkan kasus bersifat linear kemudian dapat dibedakan berdasarkan hubungan sebab akibat dan diklasifikasikan sesuai kelasnya [59]. Gambar 2.3 menunjukkan contoh penerapan *hyperplane* terbaik untuk memisahkan dua kelas.



Gambar 2. 3 *Hyperplane* Terbaik yang Memisahkan Kedua Kelas -1 Dan +1

(Sumber : Penggunaan Feature Selection di Algoritma *Support Vector Machine* untuk Sentimen Analisis Komisi Pemilihan Umum [58])

Ada banyak jenis kernel yang dimiliki oleh algoritma SVM, diantaranya kernel linear*(linear*), kernel polynomial(*non-linear*), kernel RBF gaussian(*non-linear*), dan kernel siqmoid(*non-linear)*. Rumus persamaan

yang biasa digunakan untuk keempat kernel tersebut dapat dilihat pada Tabel 2.2[60].

Tabel 2. 2 Tabel Persamaan Umum Kernel pada SVM

Nama Kernel	Rumus Persamaan
Linear	$K(x.x^i) = x.x^i$
Polinomial	$K(x.x^i) = (x.x^i + d)^i$
RBF Gaussian	$K(x.x^i) = exp(-\gamma\|\|x - x^i\|\|^2)$
Siqmoid	$K(x.x^i) = tanh(ax.x^i + \beta)$

Tahapan yang dilakukan dalam algoritma SVM adalah sebagai berikut [61][62]:

a. Menentukan variabel yang akan dijadikan kelas atau kategori.

b. Inisialisasi awal ditentukan untuk nilai α, C, λ, γ, dan epsilon.

c. Menghitung matriks dengan menggunakan persamaan (2.5)

$$D_{ij} = y_i y_j (K(\overrightarrow{x_i} \cdot \overrightarrow{x_j}) + \lambda^2) \tag{2.5}$$

Keterangan:

D_{ij} = Elemen dari matriks data ke-ij

y_i = Kategori data ke-i

y_j = Kategori data ke-j

λ = Turunan fungsi

$K(\overrightarrow{x_i} \cdot \overrightarrow{x_j})$ = Fungsi dari kernel

d. Untuk data ke n = 1,2,3 dan seterusnya dapat menggunakan persamaan (2.6), (2.7), dan (2.8)

$$E_i = \sum_{j=1}^{n} a_i D_{ij} \tag{2.6}$$

$$\delta a_i = \min\{\max[\gamma(1 - E_i), -a_i], C - a_i\} \tag{2.7}$$

$$a_i = a_i + \delta a_i \tag{2.8}$$

Keterangan:

E_i = Nilai *error* (kesalahan) pada data ke-i

γ = Tingkat pembelajaran yang sudah dilakukan

$max_{(i)} D_{ij}$ = Nilai maksimum pada diagonal matriks hesse

a_i = Bobot nilai setiap titik pada data

C = Nilai konstanta

e. Mencari nilai bobot vektor (w) penyimpangan bias (b) seperti pada persamaan (2.9) dan (2.10).

$$w = \sum_{i}^{N} a_i y_i x_i \tag{2.9}$$

$$b = -\frac{1}{2}[w \cdot x^{+} + w \cdot x^{-}] \tag{2.10}$$

Keterangan:

N = Jumlah data yang diuji

w = Bobot dari vektor

b = Nilai penyimpangan atau bias

$y_i \in \{-1, +1\}$ = Kelas atau kategori data

f. Pengujian pada data yang akan diuji

g. Perhitungan keputusan dapat dilakukan menggunakan persamaan (2.11)

$$h(x) = \begin{cases} +1, if\ w\ \cdot x + b \geq 0 \\ -1, if\ w\ \cdot x + b < 0 \end{cases} \tag{2.11}$$

Jika hasil keputusan bernilai lebih dari sama dengan 0 maka nilai h(x) adalah +1, dimana kelas atau kategori dalam data termasuk positif. Sedangkan apabila perhitungan fungsi keputusan mendapatkan hasil dengan nilai kurang dari 0 maka nilai h(x) adalah -1, dimana kelas atau kategori dalam data termasuk negatif. Perhitungan keputusan dapat dicari menggunakan persamaan (2.12)

$$h(x) = w \cdot x + b \tag{2.12}$$

$$h(x) = \sum_{i=1}^{m} a_i y_i K(x, x_i) + b \tag{2.12}$$

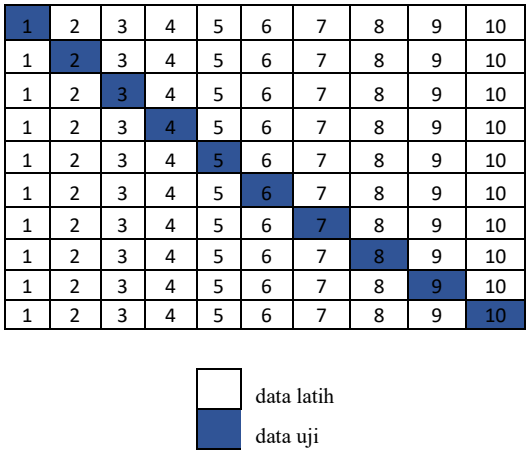
Keterangan:

m = Jumlah titik data yang memiliki nilai $a_i > 0$

h(x) = Fungsi keputusan klasifikasi sign(h(x))

2.2.6 K-Fold Cross Validation

K-Fold Cross Validation merupakan salah satu teknik untuk melakukan validasi model yang memperkirakan dan menghitung suatu keberhasilan sistem dengan cara memecah data secara random menjadi k set bagian data dengan besaran yang sama kemudian diuji sebanyak k kali[63][3]. Gambar 2.4 adalah contoh penggunaan *K-fold Cross Validation* dengan nilai k=10.



Gambar 2. 4 Model *K-Fold Cross Validation* dengan K=10

(Sumber : Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation [63])

Model penelitian seperti Gambar 2. 4 dapat diasumsikan dengan nama setiap data yaitu D1, D2, D3, D4, D5, D6, D7, D8, D9, dan D10. Percobaan dapat dilakukan sebagai berikut:

a. Percobaan pertama, data D1 digunakan sebagai data uji. Sedangkan data D2, D3, D4, D5, D6, D7, D8, D9, dan D10 digunakan sebagai data latih.
b. Percobaan kedua, data D2 digunakan sebagai data uji. Sedangkan data D1, D3, D4, D5, D6, D7, D8, D9, dan D10 digunakan sebagai data latih.
c. Percobaan ketiga, data D3 digunakan sebagai data uji. Sedangkan data D1, D2, D4, D5, D6, D7, D8, D9, dan D10 digunakan sebagai data latih.
d. Percobaan ketiga, data D4 digunakan sebagai data uji. Sedangkan data D1, D2, D3, D5, D6, D7, D8, D9, dan D10 digunakan sebagai data latih.
e. Percobaan ketiga, data D5 digunakan sebagai data uji. Sedangkan data D1, D2, D3, D4, D6, D7, D8, D9, dan D10 digunakan sebagai data latih.
f. Percobaan ketiga, data D6 digunakan sebagai data uji. Sedangkan data D1, D2, D3, D4, D5, D7, D8, D9, dan D10 digunakan sebagai data latih.
g. Percobaan ketiga, data D7 digunakan sebagai data uji. Sedangkan data D1, D2, D3, D4, D5, D6, D8, D9, dan D10 digunakan sebagai data latih.
h. Percobaan ketiga, data D8 digunakan sebagai data uji. Sedangkan data D1, D2, D3, D4, D5, D6, D7, D9, dan D10 digunakan sebagai data latih.
i. Percobaan ketiga, data D9 digunakan sebagai data uji. Sedangkan data D1, D2, D3, D4, D5, D6, D7, D8, dan D10 digunakan sebagai data latih.

j. Percobaan ketiga, data D10 digunakan sebagai data uji. Sedangkan data D1, D2, D3, D4, D5, D6, D7, D8, dan D9 digunakan sebagai data latih.

2.2.7 Akurasi

Evaluasi merupakan tahapan untuk mengukur performa dari suatu model dari algoritma yang telah dibuat menggunakan confusion matrix dengan mempertimbangkan nilai akurasi, *standard deviation, f1 score, recall, precision, sensitifity, specificity*, dan *Root Mean Square Error* (RMSE), serta *Mean Square Error* (MSE) [64] [65]. *Confusion matrix* sendiri merupakan salah satu alat dengan fungsi untuk menganalisis seberapa bisa suatu model algoritma klasifikasi dapat mengenali *tuple* dari data yang berbeda [64]. Gambar 2.5 menunjukkan parameter pada *confusion matrix*.

	Positive	Negative	
Positive	**TP**	**FN**	TP + FN
Negative	**FP**	**TN**	FP + TN
	TP + FP	FN + TN	

Gambar 2. 5 Confusion Matrix

(Sumber : Perbandingan Akurasi dan Waktu Proses Algoritma K-NN dan SVM dalam Analisis Sentimen Twitter [64])

Penjelasan mengenai parameter *confusion matrix* beserta persamaan yang digunakan adalah sebagai berikut [3][66]:

a. *True Positive* (TP) merupakan data *true* yang berhasil diklasifikasikan oleh model sebagai data *true*. Persamaan TP dapat dilihat pada persamaan (2.13).

$$TPR = \frac{TP}{TP+FN} \tag{2.13}$$

Keterangan:

TPR = *True Positive Rate*

TP = *True Positive*

FN = *False Negative*

b. *False Positive* (FP) merupakan data true yang berhasil diklasifikasikan oleh model sebagai data *false*. Persamaan FP dapat dilihat pada persamaan (2.14).

$$FPR = \frac{FP}{FP + TN} \tag{2.14}$$

Keterangan:

FPR = *False Positive Rate*

TN = *True Negative*

c. *False Negative* (FN) merupakan data *false* yang berhasil diklasifikasikan oleh model sebagai data *true*. Persamaan FN dapat dilihat pada persamaan (2.15).

$$FNR = \frac{FN}{FN + TP} \tag{2.15}$$

Keterangan:

FNR = *False Negative Rate*

FN = *False Negative*

d. *True Negative* (TP) merupakan data *false* yang berhasil diklasifikasikan oleh model sebagai data *false*. Persamaan TP dapat dilihat pada persamaan (2.16).

$$TNR = \frac{TN}{TN + FP} \tag{2.16}$$

Keterangan:

TNR = *True Negative Rate*

TN = *True Negative*

Akurasi merupakan prediksi *true* atau tingkat dari nilai kebenaran yang dihasilkan dari model algoritma dalam melakukan klasifikasi, dimana perhitungan dari akurasi dapat dilihat pada persamaan (2.17)[3][64].

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{2.17}$$

Precision merupakan tingkat ketepatan dari suatu model algoritma yang dapat dihitung dengan membandingkan nilai *True Positive* dengan jumlah label data yang bernilai *positive*, dapat dilihat pada persamaan (2.18)[64].

$$Precision = \frac{TP}{TP+FP} \tag{2.18}$$

Recall merupakan salah satu pengukuran kelengkapan dari suatu model algoritma yang dapat dihitung dengan membandingkan total data yang benar-benar bernilai *positive*, dapat dilihat pada persamaan (2.19)[64].

$$Recall = \frac{TP}{TP+FN} \tag{2.19}$$

Mengukur kesalahan dalam proses klasifikasi dapat dilakukan dengan menggunakan RMSE dan MSE, dimana apabila nilai RMSE dan MSE semakin kecil maka akurasi yang didapatkan akan semakin tinggi [67]. Persamaan untuk menghitung nilai RMSE dan MSE dapat dilihat pada persamaan (2.20) dan (2.21) [67]. (2.23)

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_1-\hat{y}_1)^2}{n}} \tag{2.20}$$

$$MSE = \frac{1}{2}\sum_{i=1}^{n}(y_1 - \hat{y}_1)^2 \tag{2.21}$$

Keterangan:

$\hat{y}_1, \hat{y}_2, \ldots \ldots, \hat{y}_n$ = nilai yang diprediksi

$y_1, y_2, \ldots \ldots \ldots \ldots, y_n$ = nilai yang diamati

Contoh perhitungan RMSE dan MSE

Positif = 1, Negatif = 0

Actual Negatif	**202**	**0**
Actual Negative	**16**	**0**
	Predict Negative	*Predict Positive*

$$RMSE = \sqrt{\frac{\begin{matrix}(1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2 \\ +(1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2+(1-0)^2 \\ +(1-0)^2+(1-0)^2\end{matrix}}{202}}$$

$$RMSE = \sqrt{\frac{16}{202}} = 0{,}2701851217221259$$

$$MSE = RMSE^2 = 0{,}073$$