BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian tentang klasifikasi penyakit kanker serviks yang telah dilakukan oleh beberapa peneliti sebelumnya dan menjadi latar belakang penelitian ini dijabarkan pada tabel dibawah ini:

A. Penelitian dengan metode klasifikasi datamining

Tabel 2.1 . Review Jurnal Klasifikasi penyakit kanker serviks dengan Berbagai Algoritma

NO	JUDUL /PENULIS /TAHUN	METODE	HASIL	KEKURANGAN
1	Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer Asadi F.1* , Salehnasab C.2, Ajori L.3 2020	SVM, QUEST, C&R Tree, MLP-ANNs dan RBF-ANNs	Akurasi SVM 93.33, Akurasi QUEST 95,55, Akurasi C&R Tree 95,55, Akurasi MLP-ANNs 90,90 dan Akurasi RBF-ANNs 95.45	Hasil Akurasi optimal
2	Supervised deep learning embeddings for the prediction of cervical cancer diagnosis, Kelwin Fernandes1,2, Davide Chicco3, Jaime S. Cardoso1,2 and Jessica Fernandes4, 2018	SVM KNN DECISION TREE	hasil prediksi yang akurat (area teratas di bawah kurva AUC = 0,6875) yang mengungguli metode yang dikembangkan sebelumnya,	masih minim langkah pemilihan fitur, yang dapat menyatakan fitur yang paling relevan di antara kumpulan data.

3	Analisa Metode Random Forest Tree dan K-Nearest Neighbor dalam Mendeteksi Kanker Serviks, Andriana1, Steelea2, Edward Suwandy Salima3, Hartato Bindana4, Endy Pranotoa5, *Abdi Dharmaa6, 2020	Random Fores, KNN	Akurasi akhir ditunjukkan 88,7% untuk Random Forest dan **90,6% untuk KNN.**	Bisa di uji kembali menggunakan metode selain KNN dan Random Forest
4	Penerapan algoritma K-Medoids untuk Pengelompokan Penyakit di Pekanbaru Riau, Tri Juninda, Mustakim, Elvia Andri, 2019	K-Medoids	Klaster 1 : 420 Klaster 2 : 349 Klaster 3 : 794 Klaster 4 : 1248	Semakin Banyak jumlah dataset semakin baik untuk memulai penelitian
5	Penerapan Metode K-Medoids Clustering Pada Penanganan Kasus Diare Di Indonesia, Fitri Hardiyanti1 , Heru Satria Tambunan2 , Ilham Syaputra Saragih3, 2019	K-Medoids	cluster 1 (rendah) = 31 cluster 2 (tinggi) =3 Jumlah 34	Cluster 2 (C2) merupakan jumlah penanganan kasus diare yang termasuk Cluster tertinggi, maka diperlukan penganan yg maksimal
6	Deteksi Pola Pasien Kanker Serviks dengan algoritma Extra Trees dan K-Nearest Neighbor, Abdi Dharmaa1, Porman Manalu a2, Gidion Stepen Sinaga a3, Riael Siringoringo a4, Imam S. Palangai a5, Kiki Setiawan a6, Andrian a7	Extra Trees, KNN	akurasi akhir Extra Trees 88% dan **KNN 89%.**	Optimisasi pengelompokan data dan parameter juga dapat dilakukan untuk penelitian selanjutnya untuk pengembangan terhadap model

7	Cervical Cancer Disease Prediction System Using CART, Naive Bayes, and k-NN, Tutus Praningki*1, Indra Budi2, 2017	CART, Naive Bayes, and k-NN,	CART 88,89%,, Naive Bayes 94,44%, , dan **k-NN 85,04%.,**	Pada penelitian ini tidak dilakukan deteksi dan remove data outlier, karena keterbatasan jumlah dataset. Dengan dilakukan remove data outlier dapat meningkatkan hasil klasifikasi.
8	Multi-Label Classification of Research Papers Using Multi-Label K-Nearest Neighbour Algorithm, Shurui Li1, a, † and Jiechen Ou2, b, † 2021	ML-KNN Decision Tree: Extra Tree KNN	ML-KNN : 0.582 Decision Tree: 0.448 Extra Tree 0.432 **KNN 0.456** **Akurasi 93%**	-

Dari hasil review beberapa jurnal baik jurnal nasional dan internasional saya menyimpulkan agar data set dapat terakurasi dengan baik maka data yang dibutuhkan peneliti semakin banyak akan semakin baik akurasinya dengan jumlah data diatas 500 fitur, dan metode yang digunakan untuk beberapa jurnal yang telah direview tingkat akurasi cenderung lebih tinggi dengan menggunakan KNN dengan akurasi hampir mencapai 91 % sesuai pada tabel berikut:

Tabel 2.2 Hasil Akurasi Review Jurnal

No	Judul	Metode Terbaik dan
1	Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer	QUEST dengan tingkat akurasi 95,55 %
2	Supervised deep learning embeddings for the prediction of cervical cancer diagnosis,	kurva AUC = 0,6875
3	Analisa Metode Random Forest Tree dan K-Nearest Neighbor dalam Mendeteksi Kanker Serviks,	KNN = 90,6 %
4	Penerapan algoritma K-Medoids untuk Pengelompokan Penyakit di Pekanbaru Riau,	K-Medoids , Klaster 4 : 1248
5	Penerapan Metode K-Medoids Clustering Pada Penanganan Kasus Diare Di Indonesia,	K-Medoids , cluster 2 (tinggi) =3
6	Deteksi Pola Pasien Kanker Serviks dengan ALGORITME Extra Trees dan K-Nearest Neighbor,	KNN dengan tingkat akurasi 89%.
7	Cervical Cancer Disease Prediction System Using CART, Naive Bayes, and k-NN,	Naive Bayes dengan tingkat akurasi 94,44%,
8	Multi-Label Classification of Research Papers Using Multi-	ML-KNN : 0.582, akurasi KNN 93%

B. Penelitian yang menggunakan metode optimasi

Tabel 2.2 Review Jurnal yang Menggunakan Forward dan lain-lain

NO	JUDUL /PENULIS /TAHUN	METODE	HASIL	KELEBIHAN
1	metode support vector machine dan Forward Selection selection prediksi pembayaran pembelian bahan baku kopra, vo Colanus Rally Drajana ivocolanusrally@gmail.com	support vector machine dan Forward Selection selection	Nilai Variabel 2, Periode RMSE 0,269.	memiliki kelebihan dalam menunjukan performa yang sangat baik untuk prediksi time series (SVM). mengoptimaikan variabel yang akan dimasukan kedalam model (Forward Selection)
2	optimasi algoritma naïve bayes menggunakan Forward Selection selection untuk klasifikasi penyakit ginjal kronis, Miftahul Rizal1 , Muhamad Zakhy Syahaf2 , Satrio Rully Priyambodo3 , Yudi Ramdhani4	C4. 5 = 90, 45%, KNN = 91, 50%, NB=92, 92%, Algoritma Logistic Regresion = 80, 09%	C4. 5 = 90, 45%, **KNN = 91, 50%,** NB=92, 92%, Algoritma Logistic Regresion = 80, 09%	penggunaan algoritma klasifikasi optimasi fiture berpengaruh signifikan terhadap peningkatan akurasi. Ber-dasarkan hasil penelitian yang diperoleh pada dataset penyakit ginjal kronis dapat diketahui bahwa optimasi algoritma Naïve Bayes menggunakan Forward Selection selection dapat meningkatkan hasil akurasi
3	Algoritma naïve bayes berbasis Forward Selection selection untuk Prediksi bimbingan konseling siswa, M. Rudi Fanan 2020	Naïve Bayes, Forward Selection Selection	Naïve Bayes = 94.55% Naïve Bayes +Forward Selection = u 94.84%	pengujian data bimbingan dan konseling siswa menggunakan metode Naïve Bayes berbasis Forward Selection Selection terbukti dapat meningkatkan tingkat akurasi dengan adanya penggunaan fitur seleksi tersebut, dibandingkan dengan metode Naïve Bayes tunggal, yang ditandai dengan peningkatan nilai akurasi sebesar 0.29%
4	metode k-nearest neighbor berbasis Forward Selection selection untuk prediksi harga komoditi lada. Muis Nanja1, Purwanto2, 2015	1.KNN, 2.KNN+ Forward Selection 3.KNN+ Bacward 4.SVM+ Forward Selection 5.SVM+	KNN+Forward Selection = 1559,741 RMSE	fitur seleksi yakni Forward Selection selection merupakan model yang lebih baik dalam melakukan seleksi variabelyang signifikan dibandingkan dengan bacward elimination.Algoritma KNN berbasis Forward Selection selectiontelah menunjukan performa yang lebih baik dibandingkan algoritma KNN, KNN berbasis backward eliminationdan SVM

2.2 Data Mining

Penambangan data atau yang dikenal sebagai Data Mining merupakan proses analisis data dari berbagai perspektif dengan tujuan untuk mengidentifikasi informasi yang berharga. Informasi ini dapat digunakan untuk meningkatkan keuntungan, mengurangi biaya transaksi, atau bahkan mencapai keduanya. Secara teknis, penambangan data dapat dijelaskan sebagai metode untuk menemukan korelasi atau pola dalam data yang terdapat dalam basis data dunia nyata yang besar. Kemampuan data mining untuk mengungkap informasi dari basis data yang luas sering diibaratkan seberti menambang emas di wilayah yang kaya. Teknologi ini memiliki berbagai aplikasi termasuk:

a) Prediksi tren dan karakteristik bisnis, dengan data mining yang otomatis mencari informasi prediktif dalam basis data besar
b) Mengungkap pola-pola yang belum pernah di ketahui sebelumnya, melalui data mining yang secara menyeluruh mengeksplorasi basis data dan mengidentifikasi pola-pola yang sebelumnya tersembunyi.

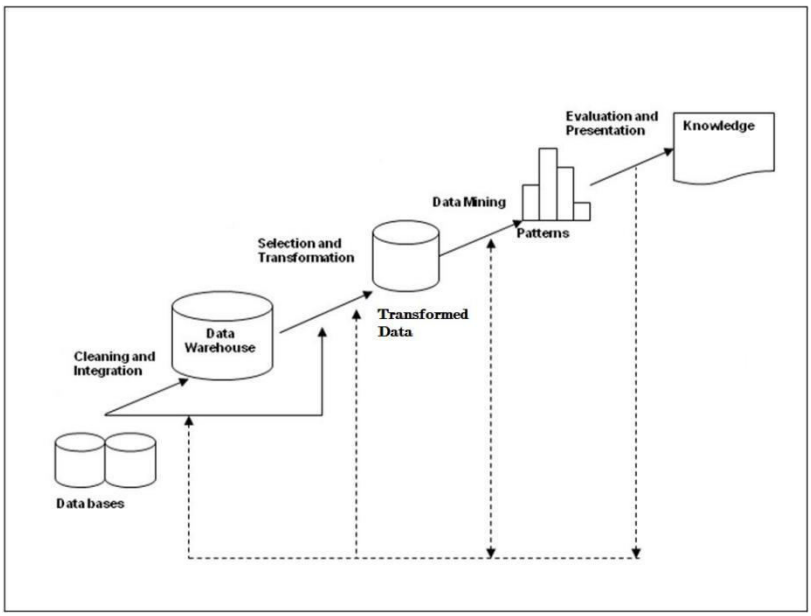
Secara teknis, data mining dapat di jelaskan sebagai pendekatan yang digunakan untuk mengidentifikasi sebagian kecil dari data yang sangat besar. Proses pengumpulan dan analisis data seperti itu dapat diimplementasikan melalui perangkat lunak komputasi statistik, komputasi matematis, atau bahkan kecerdasan buatan (AI). Dalam data mining, ada lima peran utama, yaitu estimasi, prediksi, pengidentifikasian asosiasi, klasifikasi, dan pengelompokan data.

Data mining adalah rangkaian Langkah untuk mengeksplorasi nilai tambah berupa informasi yang sebelumnya tidak diketahui secara manual dari sekumpulan data[15]. Data Mining memiliki sejumlah keunggulan sebagai alat analisis diantaranya:

a) Mampu mengelola data dalam jumlah besar dan dengan tingkat kompleksitas yang tinggi
b) Dapat menangani beragam jenis atribut data
c) Kemampuan dalam mencari dan memproses data secara otomatis, meskipun dalam beberapa teknik data mining, beberapa parameter perlu diatur secara manual oleh pengguna.
d) Mampu memanfaatkan pengalaman serta kesalahan sebelumnya untuk meningkatkan kualitas dan hasil analisis dan menghasilkan hasil yang optimal

Sedangkan data mining memiliki kekurangan dalam pencarian data, yaitu tidak melakukan pencarian secara individu, melainkan mengelompokan data berdasarkan kriteria tertentu.

Data mining adalah satu tahap dalam Knowledge Discovery in Database (KDD).



Gambar 0.1 Proses KDD

[16]. Selama proses penambangan data ini, dapat melakukan klasifikasi, prediksi, dan pengkategorian untuk mengungkap informasi yang bermanfaat dalam kumpulan data yang sangat besar. [17]. Penggunaan istilah "Data Mining" dan "KDD" seringkali digunakan secara bergantian untuk menjelaskan proses yang lebih lanjut dalam menggali informasi yang tersembunyi dalam basis data besar. [18]

Berikut adalah uraian dari tahapan proses KDD yang tergambar pada gambar 2.1

2.2.1 Data Selection

Sebelum memulai proses penambangan data, langkah awal adalah menyelesaikan proses pemilihan data dari seluruh Kumpulan data operasional. Pada tahap ini, keputusan mengenai pemilihan dataset atau fokus pada subset variabel tertentu

sangat penting, karena proses ini memerlukan waktu. Hasil seleksi data yang akan digunakan dalam proses penambangan data akan disimpan dalam dokumen terpisah yang berbeda dari basis data operasional.

2.2.2 Preprocessing atau cleaning

Setelah menyelesaikan tahap pemilihan data, Langkah berikutnya dalam data mining adalah *preprocessing* data. Tahap ini melibatkan proses pembersihkan data untuk mengidentifikasi dan menghapus data ganda, data dengan kesalahan sintaksis dan memaksimalkan hasil analisis. *Preprocessing* data merupakan tahap yang sangat penting dalam proses data mining, karena kualitas hasil data mining sangat dipengaruhi oleh proses *preprocessing* data.

2.2.3 Transformasi

Transformasi data adaah Langkah yang mengacu pada mengubah bentuk dari satu bentuk ke dalam bentuk lainnya. Tujuan utama dari transformasi data adalah untuk meningkatkan kualitas data dan mengoptimalkan kinerja algoritma pemrosesan data. Beberapa contoh umum dari transformasi data meliputi:

a) Normalisasi: Proses mengubah data kedalam skala yang seragam, sehingga memungkinkan perbandingan dan analisis yang lebih mudah.
b) Penghapusan duplikat: proses menghilangkan data yang sama dalam satu data set
c) Imputasi: proses mengisi nilai yang hilang dalam data dengan nilai yang sesuai
d) Pemfilteran: proses pemilihan subset data berdasarkan kriteria tertentu.

e) Encoding: proses mengubah format data dari suatu bentuk ke bentuk lain, seperti mengonversi data nominal menjadi data numerik
f) Reduksi dimensi: proses mengurangi jumlah fitur atau variabel dalam data

Transformasi data memiliki peran sangat penting dalam pemrosesan data, karena dapat membantu meningkatkan kualitas data dan mengoptimalkan kecepatan pemrosesan data.

2.2.4 Proses Penambangan Data

Proses penambangan data adalah tahap dimana pola atau informasi yang sebelumnya tidak diketahui dapat diidentifikasi dengan metode khusus. Berbagai Teknik, metode, dan algoritma yang digunakan dalam proses ini sangat bervariasi. Pemilihan metode dan algoritma bergantung pada tujuan dan konteks keseluruhan penambangan data. Proses ini seringkali memanfaatkan metode matematika, statistika dan juga kecerdasan buatan.

2.2.5 Interprestation atau Evaluation

Hasil dari proses penambangan data harus disajikan dalam format yang mudah dimengerti oleh berbagai pemangku kepentingan yang tertarik dengan hasil penambangan data tersebut. Pada tahap ini, dilakukan pemeriksaan pola informasi berdasarkan fakta-fakta yang telah ada sebelumnya.

2.3 Klasifikasi

Klasifikasi adalah Salah satu fitur dalam data mining yang melibatkan pembuatan model untuk meramalkan atau mengelompokan objek-objek dalam basis data ke dalam kelas atau kategori tertentu [19]. Klasfikasi adalah Tindakan mengelompokan informasi berdasarkan atribut-atribut yang diberikan label,

sehingga algoritma-algoritma yang digunakan untuk menangani permasalahan klasifikasi termasuk dalam kategori *administered kearning* atau pembelajaran yang di awasi. Hasil dari proses ini adalah mencapai tingkat akurasi atau presisi yang spesifik.

2.4 Teori Dasar

a) Fitur Seleksi

Fitur seleksi adalah proses memilih subset fitur yang paling relevan dan informatif dari dataset. Proses ini bertujuan untuk:

- Meningkatkan kinerja model pembelajaran mesin: Dengan memilih fitur yang relevan, model dapat belajar lebih efektif dan menghasilkan prediksi yang lebih akurat.
- Meningkatkan efisiensi: Dengan mengurangi jumlah fitur, waktu dan sumber daya yang dibutuhkan untuk melatih dan menjalankan model dapat dikurangi.
- Meningkatkan interpretabilitas: Dengan memilih fitur yang mudah dipahami, model menjadi lebih mudah dipahami dan diinterpretasikan.

Ada beberapa metode yang dapat digunakan untuk melakukan fitur seleksi:

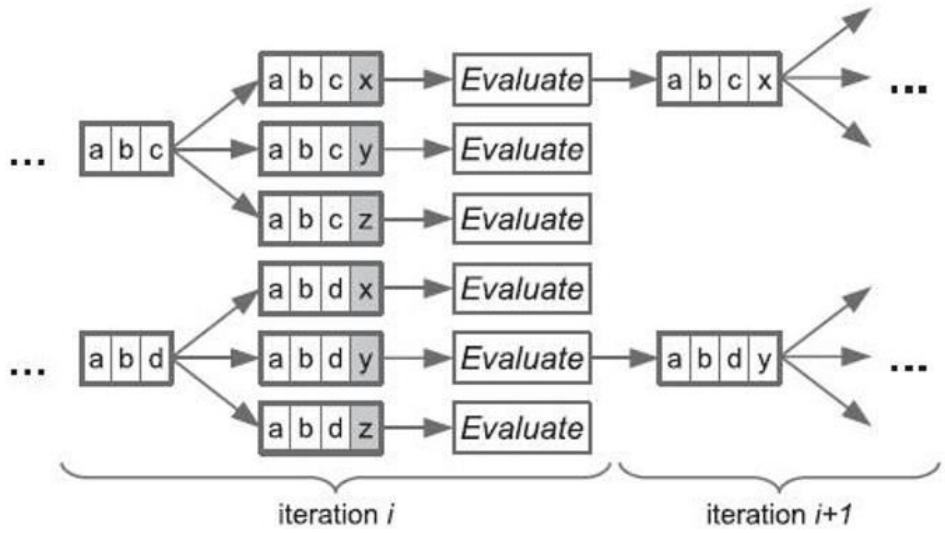
- Metode filter: Metode ini mengevaluasi setiap fitur secara individual dan memilih fitur berdasarkan skor yang dihitung. Contoh metode filter adalah:
    - *Information gain*: Mengukur seberapa banyak informasi yang diberikan oleh suatu fitur tentang target variabel.

  - *Chi-squared test*: Mengukur hubungan statistik antara suatu fitur dan target variabel.
- Metode wrapper: Metode ini mengevaluasi subset fitur secara keseluruhan dan memilih subset yang menghasilkan kinerja terbaik pada model pembelajaran mesin. Contoh metode wrapper adalah:
  - *Sequential Forward Selection* : Memulai dengan subset kosong dan menambahkan fitur secara berurutan yang paling meningkatkan kinerja model.
  - *Sequential backward selection*: Memulai dengan semua fitur dan menghapus fitur secara berurutan yang paling sedikit menurunkan kinerja model.
- Metode embedded: Metode ini mengintegrasikan proses seleksi fitur dengan proses pelatihan model. Contoh metode embedded adalah:
  - Lasso regression: Memilih fitur dengan menambahkan penalti pada koefisien regresi yang besar.
  - Tree-based methods: Memilih fitur berdasarkan importance score yang dihitung oleh model.

Pemilihan metode fitur seleksi yang tepat tergantung pada beberapa faktor:

- Jenis data: Apakah datanya numerik, kategorikal, atau campuran?
- Ukuran dataset: Apakah datasetnya kecil, besar, atau sangat besar?
- Ketersediaan alat dan sumber daya: Apakah Anda memiliki alat dan sumber daya yang diperlukan untuk menjalankan metode fitur seleksi yang dipilih?

Forward Selection selection juga dikenal merupakan prosedur penting dalam persiapan data dalam penambangan data untuk mengoptimalkan kinerja dan mempercepat eksekusi suatu algoritma. Teknik memungkinkan pemilihan yang luas dari subset fitur dengan jumlah pilihan yang banyak, sehingga mengucualikan fitur yang tidak penting dan relevan untuk klasifikasi. Forward Selection selection juga dikenal sebagai pemilihan fitur, pemilihan subset, pemilihan atribut, atau pemilihan variabel, juga dapat didefinisikan sebagai proses pemilihan atribut yang sesuai untuk digunakan dalam proses klasifikasi atau pengelompokan untuk tujuan pengecualian. Kompleksitas algoritma klas-ifikasi, menyetel akurasi algoritma klasifikasi, dan menentukan atribut yang mempengaruhi tingkat akurasi [13].



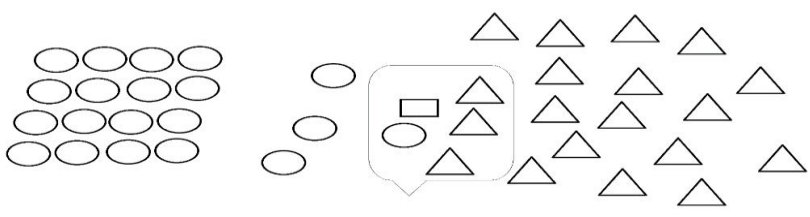
Gambar 1. Metode *Forward Selection*

b) K-NN

K-Nearest Neighbor atau KNN adalah suatu algoritma yang memangaatkan data latih (training record) yang diperoleh dari tetangga terdekat (nearest neighbor) dalam melakukan klasifikasi data, dengan jumlah tetangga terdekat yang di tentukan oleh nilai K [22]. Algoritma KNN mengikuti serangkaian tahap kerja yang meliputi:

1) Langkah pertama adalah menentukan jumlah tetangga yang ingin digunakan
2) Selanjutnya menghitung jarak antara data baru dengan seluruh tetangga yang ada, menggunakan rumus perhitungan seperti Euclidean Distance atau metode lainnya
3) Setelah itu, memilih k tetangga terdekat dengan jarak terkecil untuk digunakan dalam proses pengambilan keputusan prediksi berdasarkan hasil perhitungan jarak yang telah dilakukan sebelumnya.

K-Nearest Neighbor (KNN) adalah suatu algoritma pembelajaran mesin yang berfokus pada pembelajaran non-parametrik dan konsep Lazy Learning. Berikut ini adalah gambar yang di temukan dalam buku "Data Mining untuk Klasifikasi Data" yang ditulis oleh Dr. Suyanto mengenai konsep dasar KNN:



Gambar 0.2 Konsep dasar KNN

Tujuan algoritma ini adalah untuk melakukan klasifikasi pada objek-objek baru berdasarkan atribut-atribut dan data latih. Pada tahun 1968, Cover dan Hart memperkenalkan algoritma K-Nearest Neighbor (KNN) sebagai algoritma yang dikenal sebagai algoritma "malas" karena pendekatannya melibatkan penyimpanan semua data pelatihan dan pembentukan model yang ditunda hingga data uji diberikan kepada algoritma ini untuk diprediksi. [23]. Rumus yang mendasari proses KNN adalah sebagai berikut:

$$distance = \sqrt{\sum_{i=1}^{n}(x_{training}^{i} - x_{testing})^2} \quad \text{..... (2)}$$

Keterangan:

$x_{training}^{i}$: data training ke-i

$x_{testing}$: data testing

i : record data ke-i dari tabel

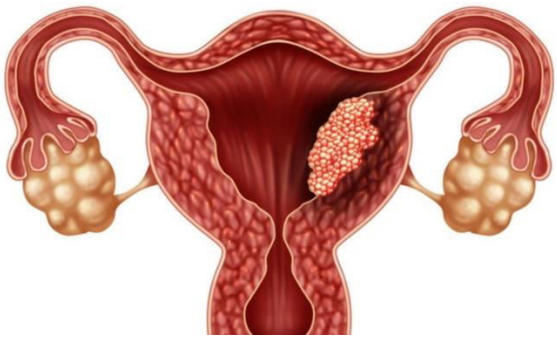
n : jumlah data training

secara sederhana, proses K-NN dapat diibaratkan sebagai sistem pemungutan suara dalam klasifikasi data. Nilai K dalam K-NN dapat diperoleh dengan mengukur jarak, kemiripan atau ketidakmiripan antara data. Ketika melakukan klasifikasi, K-NN akan memeriksa semua pola yang ada dalam data latih untuk menemukan pola klasifikasi yang sesuai. Hal yang menarik, algoritma K-NN

beroperasi secara lokal, hanya mempertimbangkan nilai K terdekat, sehingga sesuai dengan kelompok data yang terdekat.

2.5 Kanker Serviks

Kanker leher Rahim merujuk pada kondisi kanker yang muncuk di serviks uterus, yang merupakan organ reproduksi Wanita berada di antara Rahim (uterus) dan Liang senggama yang berfungsi sebagai pintu masuk menuju Rahim. Beberapa factor yang dapat menjadi penyebab kanker leher Rahim diantaranya menikah pada usia muda, paparan Human Papilloma Virus (HPV), kurangnya kebersihan genitalia, kebiasaan merokok, Riwayat penyakit kulit kelamin seperti herpes dan kutil genital, frekuensi kehamilan yang tinggi. Faktor-faktor lainnnya melibatkan trauma kronis pada serviks seperti saat persalinan, infeksi, iritasi berkepanjangan, paparan mikroba, radiasi, atau kontaminasi oleh bahan kimia. Penggunaan antiseptic secara rutin juga dapat menyebabkan iritasi pada rahim dan memicu perkembangan kanker.



Gambar 0.3 Kanker Serviks

Pada stadium awal, kanker jenis ini cenderung tidak terdeteksi bahkan penderitanya tidak sadar bila dirinya sudah terkena kanker serviks ini. Namun jika

kanker sudah berkembang dan semakin memburuk, maka akan timbul indikasi gejala seperti keputihan yang semakin lama semakin berbau busuk, berwarna kekuningan kental, pendarahan setelah melakukan hubungan seksual, dan bahkan apabila sudah sangat parah, akan terjadi pendarahan spontan walau tidak melakukan hubungan seksual.

2.6 Rapid Miner

Dalam penelitian ini, penulis memutuskan untuk menggunakan aplikasi Rapid Miner. Alasan pemilihan Rapid Miner adalah karena kemampuannya untuk menggabungkan visualisasi, statistik, dan informasi penting tentang model kedalam laporan dengan cepat dan mudah. Rapid Miner juga menyediakan pelaporan yang informatif, yang memungkinkan pengguna untuk melihat Riwayat alur kerja untuk setiap elemen dan visualisasi langsung dari laporan tersebut. Kelebihan lainnya adalah antarmuka pengguna yang intuitif, yang memungkinkan pengguna untuk fokus pada analisis data daripada harus terjebak dalam pengkodean yang rumit. Rapid miner juga dapat membuat pembuatan *pipeline* analisis data yang kompleks menjadi lebih sederhana. RapidMiner adalah sebuah aplikasi yang popular digunakan untuk melakukan pemrosesan data mining. Aplikasi ini terkenal karena menyediakan beragam algoritma untuk tugas seperti klasifikasi, pengelompokan, dan analisis regresi dalam proses penambangan data. RapidMiner dirancang dengan antarmuka pengguna yang ramah, terdiri dari tiga perspektif utama yakni *welcome*, *design* dan *result* [25]. Pengguna dapat memulai dengan mengimpor data kedalam aplikasi, kemudian data tersebut dapat dihubungkan dengan operator, termasuk operator untuk mengatasi *missing values*.

Data yang lengkap akan dimodelkan dengan menggunakan tampiland esign yang terhubung dengan operator untuk menjalankan proses penambangan data.