BAB II
LANDASAN TEORI

2.1 Penyakit Kanker Payudara

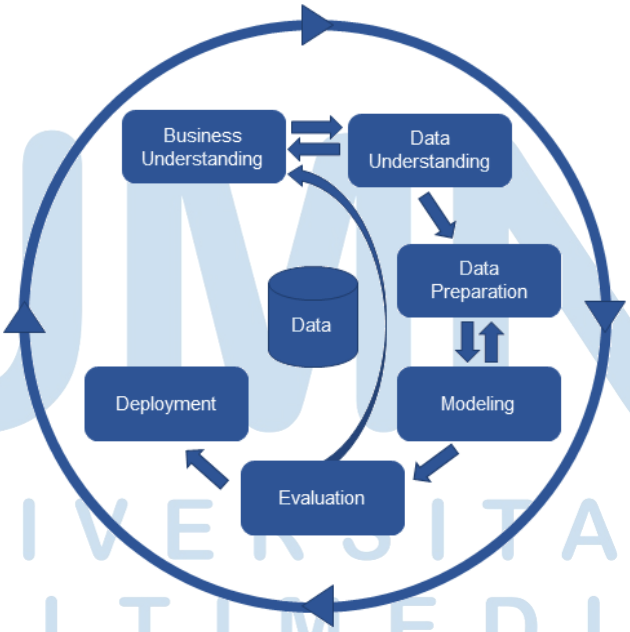
Penyakit kanker payudara yaitu kanker yang paling umum ditemukan pada wanita yang terbentuk di sel-sel payudara. Kanker payudara biasanya terasa seperti benjolan yang akan terlihat apabila dilakukan *x-ray*. Kanker payudara tidak selalu ganas tergantung letak pertumbuhan tumornya, namun kanker ini dapat menyebar ketika sel kanker masuk ke darah atau ke getah bening dan terbawa ke bagian tubuh lainnya [14]. Namun kanker payudara ini dapat mengancam kesehatan fisik maupun mental wanita di dunia. Penderita kanker payudara stadium awal cenderung memiliki tingkat kelangsungan hidup yang lebih tinggi daripada mereka yang di stadium menengah dan akhir. Karena itulah penting dilakukan deteksi dan pengobatan dini agar dapat cepat mendeteksi dan melakukan pengobatan sehingga kelangsungan hidup penderita kanker menjadi lebih tinggi [15]. Saat ini terdapat sekitar 17 macam jenis kanker payudara mulai dari yang bersifat *invasive*, *non-invasive*, *metastatic*, dan *molecular* [16].

Agar dapat mengidentifikasi kanker payudara secara cepat dan akurat dapat dilakukan pemeriksaan menggunakan *imaging techniques. Imaging techniques* ini mencakup pemeriksaan *mammography* (MG), *ultrasonography* (US), *magnetic resonance imaging* (MRI), *positron emission computed tomography* (PET), *computed tomography* (CT), dan *single-photon emission computed tomography* (SPECT). Namun teknik ini memakan biaya yang besar dan dapat membahayakan mengingat menggunakan radiasi yang tinggi [15]. Adapun alternatif lain pemeriksaan kanker payudara yaitu melalui biopsi. Biopsi adalah tindakan mengambil potongan kecil dari area tubuh dengan menggunakan jarum atau membuat sayatan kecil agar potongan kecil tersebut dapat dilakukan pemeriksaan kanker [17].

Apabila seseorang terkena kanker payudara, terdapat beberapa macam alternatif pengobatan yang dapat dilakukan. Pertama yaitu dengan melakukan operasi. Terdapat dua macam operasi payudara yaitu *breast-conserving surgery* atau *mastectomy*. *Breast-conserving surgery* dilakukan untuk mengangkat sel kanker dan beberapa jaringan normal disekitar payudara, operasi pengangkatan bergantung letak dan ukuran tumor payudara serta faktor lainnya. Sedangkan *mastectomy* yaitu tindakan pengangkatan seluruh payudara dan jaringan di dekatnya, terkadang juga dilakukan *mastectomy* ganda [18].

2.2 CRISP-DM

Cross Industry Standard Process for Data mining atau dapat disingkat sebagai CRISP-DM merupakan sebuah *framework* yang disusun sebagai langkah sederhana yang dapat membantu dalam menjalankan suatu proses *data mining*. Dalam CRISP-DM sendiri memiliki enam fase yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, dan *Deployment* [19]. Fase dari CRISP-DM akan direpresentasikan oleh gambar berikut.



Gambar 2.1 Fase CRISP-DM [20]

Setiap fase CRISP-DM tentunya memiliki peranan masing-masing. Berikut penjelasan masing-masing setiap fase dari CRISP-DM [21].

a. *Business Understanding*

   Business Understanding merupakan fase untuk memperoleh suatu masukan mengenai situasi bisnis sesuai dengan kebutuhan dan sumber daya. Misalnya dalam hal menentukan jenis proses *data mining* yang akan dilakukan serta bagaimana kriteria berhasilnya proses *data mining*.

b. *Data Understanding*

   Data Understanding merupakan fase untuk mengumpulkan data, melakukan eksplorasi data, mendeskripsikan, serta memeriksa kualitas data dari *data sources* agar data yang digunakan lebih konkret.

c. *Data Preparation*

   Data Preparation merupakan fase untuk melakukan pemilihan data, biasanya dilakukan *cleansing data*. Fase ini dilakukan agar dapat meningkatkan kualitas data agar dapat dilakukan *modelling* dengan maksimal.

d. *Modelling*

   Modelling merupakan fase untuk melakukan pemodelan pada data. Misalnya dalam melakukan pemilihan model yang digunakan, pembuatan *train case* maupun *test case*.

e. *Evaluation*

   Evaluation merupakan fase untuk melakukan pengecekan terhadap hasil yang diperoleh dari pemodelan dengan *Business Understanding.*

f. *Deployment*

   Deployment merupakan fase untuk melakukan inovasi baru berupa implementasi dari pemodelan yang telah di buat. Dalam *deployment* biasanya dilakukan proses *development, monitoring,* dan *maintenance.*

2.3 *Support Vector Machine*

Support Vector Machine atau sering disebut dengan SVM merupakan salah satu algoritma yang sering digunakan untuk melakukan klasifikasi dan regresi.

Prinsip dasar SVM yaitu untuk menemukan *hyperline* yang optimal [22]. *Hyperline* yang optimal ini nantinya akan diorientasikan sejauh mungkin dari *support vector*. Adapun *traning dataset* pada SVM akan dirumuskan sebagai berikut pada rumus 2.1.

$$(x_1, y_1), \dots, (x_n, y_n), x_i \epsilon R^d \; dan \; y_i \epsilon (-1, +1)$$

Rumus 2.1 Persamaan *Training Data* [23]

Dimana x_i merepresentasikan *feature vector* dan y_i merepresentasikan *class label* baik dalam nilai positif maupun negatif. *Hyperlane* yang optimal dapat dirumuskan kedalam rumus 2.2.

$$wx^T + b = 0$$

Rumus 2.2 Persamaan *Hyperlane* [23]

Dimana w merepresentasikan nilai *weight vector*, x merepresentasikan *input feature vector*, dan b merepresentasikan nilai bias. Adapun nilai w dan b dapat memenuhi rumus 2.3 dan 2.4.

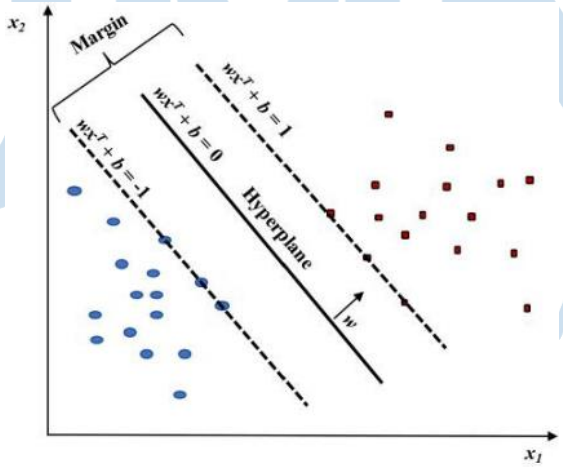
$$wx_i{}^T + b \geq +1 \; jika \; y_i = 1$$

Rumus 2.3 Persamaan *Margin* 1 [23]

$$wx_i{}^T + b \leq -1 \; jika \; y_i = -1$$

Rumus 2.4 Persamaan *Margin* 2 [23]

Training dari pemodelan SVM ini bertujuan untuk menemukan nilai w dan b untuk memaksimalkan margin seperti pada gambar 2.2 [23].



Gambar 2.2 *Hyperlane* pada SVM [23]

Selain itu, SVM juga memiliki fungsi *Kernel* untuk memetakan *classification data.* Penggunaan SVM dengan *Kernel* dilakukan sesuai dengan parameter terkait seperti *penalty parameter* (C), *Gamma parameter* (γ) yang berguna untuk mengelola tahapan pembelajaran pada algoritma SVM sehingga dapat berdampak pada hasil akurasinya [24].

Pada algortima SVM ini terdapat beberapa *kernel*. *Kernel Linear* yaitu *kernel* yang paling sederhana yang berfungsi menganalisis data yang terklasifikasi secara *linear* [25]. *Kernel Polynomial* adalah *kernel* yang bisanya cocok untuk digunakan untuk memproses *image* [25]. *Kernel RBF* yaitu *kernel* yang digunakan dalam suatu data yang biasanya tidak terpisah atau tidak terklasifikasi secara *linear* (*non-linear*) [26]. Berikut merupakan beberapa *Kernel* yang disajikan pada tabel 2.1 yang digunakan untuk mempersiapkan model SVM:

Tabel 2.1 *Kernel* yang dapat digunakan untuk mempersiapkan Model SVM [24]

KERNELS	***FORMULA***
Linear	$K(x_i, x_j) = 1 + x_i^T x_j$
Polynomial	$K(x_i, x_j) = \exp(1 + x_i^T x_j)^p$
Radial Basis Function (RBF)	$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2)$

2.4 *Hyperparameter*

Hyperparameter adalah suatu metode untuk pengklasifikasian *machine learning* dalam menentukan dan meningkatkan *performance* maupun akurasi. *Hyperparameter* juga berperan dalam mempengaruhi model pembelajaran misalnya dalam hal *machine learning.* Tak hanya itu *hyperparameter* juga berperan dalam menentukan *construction*, dan menentukan nilai evaluasi yang tepat dan terbaik dalam melakukan proses klasifikasi dalam *machine learning*. Selain itu, *hyperparameter* juga berperan dalam menemukan parameter terbaik dalam pengklasifikasian *machine learning* sehingga dapat membantu memberikan nilai akurasi sebaik mungkin [27].

2.5 *Label encoding*

Dalam melakukan *encoding*, salah satu teknik yang biasa digunakan yaitu *label encoding*. *Label encoding* merupakan suatu teknik untuk melakukan encodes dari dataset yang memiliki tipe data string ke dalam tipe data *integer*. *Label encoding* ini diterapkan dengan tujuan agar mempermudah proses *machine learning* memahami data yang digunakan dalam tahapan *Data Understanding*. Dalam proses melakukan *label encoding*, biasanya dataset yang memiliki tipe data string dilakukan encoded sehingga datanya berubah menjadi kombinasi angka saja [28].

2.6 *Confusion matrix*

Confusion matrix biasanya sering digunakan dalam *machine learning* khususnya dalam *supervised classification* atau dalam menentukan *classification* model. Struktur *confusion matrix* direpresentasikan ke dalam bentuk baris dan kolom dimana baris biasanya mengandung *class* actual dan kolom mengandung *class* predicted. Apabila direpresentasikan ke dalam binary *classification*, *confusion matrix* biasanya ditampilkan sebagai matriks 2 x 2. Dalam *confusion matrix* juga terdapat empat measures yaitu *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN) [29]. *True positive* adalah jumlah suatu data yang positif yang terklasifikasi benar oleh sistem sedangkan *true negative* adalah jumlah suatu data yang negatif yang terklasifikasi benar oleh sistem. *False positive* adalah jumlah suatu data yang negatif yang terklasifikasi benar oleh system sedangkan *false negative* adalah jumlah suatu data yang positif yang terklasifikasi salah oleh system [30]. Berikut merupakan gambar matriks dari *confusion matrix* seperti yang direpresentasikan pada gambar 2.3.

	Predicted true	Predicted false
Actual true	True positive (TP)	False negative (FN)
Actual false	False positive (FP)	True negative (TN)

Gambar 2.3 *Confusion matrix* [31]

2.7 *Classification report*

Classification report adalah performance *evaluation matrix* dalam *machine learning* yang biasanya berisi *Report* yang terdiri dari nilai *accuracy*, *precision*, *recall*, dan *f1-score*.

2.7.1 *Accuracy*

Accuracy adalah nilai *output* suatu model yang biasanya digunakan untuk merepresentasikan suatu model yang baik yang diprediksi dalam nilai *true positive* maupun *false negative* dengan dibandingkan keseluruhan data. Nilai *accuracy* yang tinggi biasanya menandakan model yang dibuat itu baik [31]. Dalam mencari nilai *accuracy*, dapat menggunakan rumus 2.5 sebagai berikut.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Rumus 2.5 Rumus Mencari *Accuracy* [31]

2.7.2 *Precision*

Precision adalah nilai yang mengandung ratio dari nilai *true positive* dengan dibandingkan suatu data yang diprediksi *positive* [31]. Untuk nilai *precision*, dapat digunakan perhitungam rumus 2.6 sebagai berikut.

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

Rumus 2.6 Rumus Mencari *Precision* [31]

2.7.3 *Recall*

Recall merupakan nilai dari total positive *classification*s dari kelas sebenarnya. Biasanya *recall* mewakili jumlah suatu data yang diprediksi *true positive* dibandingkan jumlah total seluruh data positive [31]. Dalam mencari nilai *recall* dapat digunakan rumus 2.7 berikut.

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

Rumus 2.7 Rumus Mencari *Recall* [31]

2.7.4 *F1-score*

Nilai *f1-score* adalah nilai yang mewakili gabungan antara nilai *precision* dan *recall*. *F1-score* biasanya mengandung nilai rata-rata dari kedua nilai tersebut sehingga perlu untuk dilakukan pengamatan antara *false positive* dan *false negative*. Untuk menghitung nilai *f1-score* dapat dicari dengan rumus 2.8 berikut.

$$f1 - score = 2\ x\ \frac{Precision\ x\ Recall}{Precision +\ Recall}$$

Rumus 2.8 Rumus Mencari *f1-score* [31]

2.8 *Tools*

Pada penelitian ini menggunakan *tools Jupyter Notebook. Jupyter Notebook* adalah *software* yang berbasis *open source* untuk membantu *user* dalam membuat, mengubah, menjalankan, dan membagikan *code* yang telah dibuat dengan mudah. *Jupyter Notebook* yang dikembangkan pada tahun 2014 ini dapat di *install* secara gratis serta dapat membantu dalam *interactive coding* [32], [33]. Pada *Jupyter Notebook* setiap *notebook* yang dibuat biasanya dapat disimpan dalam format ekstensi .ipynb maupun dalam format lain misalnya dalam .py untuk pemrograman *Python*. Format ini memudahkan pengguna untuk membagikan *code* yang mereka buat baik melalui *email* maupun *dropbox* [33].

2.9 Penelitian Terdahulu

Dalam memilih *framework* penelitian, digunakan beberapa penelitian terdahulu sebagai acuan. Ada dua penelitian yang digunakan dalam memilih *framework.* Pada tabel 2.2 akan ditampilkan dua penelitian yang digunakan sebagai acuan untuk memilih *framework.*

Tabel 2.2 Penelitian Terdahulu yang Menggunakan CRISP-DM

No	Nama Journal,	Judul; Penulis	Hasil Penelitian	Nilai yang Diperoleh di Penelitian

	Vol, No, Tahun			
1	Jurnal Sains dan Informatika, 07, 02, 2021 [34]	"Perbandingan Algoritma Klasifikasi untuk Prediksi Cacat Software dengan Pendekatan CRISP-DM"; Nutriana Hidayati, Joko Suntoro, Galet Guntoro Setiaji	Dari penelitian yang menggunakan kerangka kerja CRISP-DM ini diperoleh hasil bahwa teknik *machine learning* dengan algoritma *k-Nearest Neighbor (k-NN), Naïve Bayes (NB),* dan *Decision Tree* mampu memprediksi cacat *software*.	Penelitian ini menunjukkan bahwa kerangka kerja CRISP-DM efektif digunakan dalam penelitian.
2	Jurnal TECHNO Nusa Mandiri, 17, 1, 2020 [35]	*"Data Mining for Predicting The Amount of Coffee Production Using CRISP-DM Method*"; Ali Khumaidi	Penelitian yang menggunakan *framework* CRISP-DM diperoleh hasil bahwa dengan kalkulasi RMSE dengan software *RapidMiner* diperoleh prediksi dengan akurasi yang baik.	Penelitian ini diperoleh informasi bahwa dalam membuat model klasifikasi *data mining* dapat digunakan *framework* CRISP-DM.

Terdapat beberapa penelitian terdahulu yang menjadi acuan dan berhubungan dengan penelitian terkini. Pada tabel 2.3 akan ditampilkan tabel perbandingan dari penelitian terdahulu.

Tabel 2.3 Penelitian Terdahulu yang menjadi Acuan

No	Nama Journal, Vol, No, Tahun	Judul; Penulis	Hasil Penelitian	Nilai yang Diperoleh di Penelitian
1	*Procedia Computer Science*, Vol.191, 2021 [7]	*"Machine learning Algorithms for Breast Cancer Prediction and Diagnosis"*; Mohammed Amine Naji, Sanaa El Filali, Kawtar Aarika, El Habib Benlahmar, Rachida Ait Abdelouhahid, Olivier Debauche	Penggunaan algoritma *Support Vector Machine* (SVM) untuk mendeteksi kanker payudara berhasil mencapai akurasi tertinggi yaitu di angka 97,2% dengan besar nilai presisi sebesar 97,2%.	Penelitian ini memberikan gambaran tentang bagaimana algoritma SVM memprediksi kanker payudara dengan pengukuran akurasi.
2	*International Journal of Emerging Trends in Engineering Research*,	*"Predicting the Possibility of Cancer with Supervised Learning Algorithms"*;	*Support Vector Machine* (SVM) menjadi algoritma yang cocok untuk memprediksi kanker karena	Penelitian menunjukkan bagaimana memprediksi terjadinya kanker

	Vol. 8, 2020 [9]	Beena G. Pillai, I. Jeena Jecob, Madhurya J. A., A. K. Saritha	memperoleh akurasi tertinggi yaitu sebesar 97,2%.	dengan SVM secara teknis.
3	2020 *International Conference on Computing and Data Science* (CDS), Vol -, 2020 [8]	*"Diagnosis of Breast Cancer Based on Support Vector Machine and Random Forest Method";* Yuyao Wu	SVM menjadi algoritma terbaik pada penelitian ini meskipun hasil akurasinya tidak berbeda jauh yakni SVM sebesar 97% sedangkan *Random Forest* sebesar 96%.	Penelitian ini menunjukkan gambaran bahwa algoritma SVM dapat dilihat akurasinya menggunakan ROC dan F1 Score.
4	2021. *International Federation for Medical and Biological Engineering*, Vol -, 2021 [10]	*"Computational Predictions for Protein Sequences of COVID-19 Virus via Machine learning Algorithms";* Walaa Alkady, Muhammad Zanaty, Heba M. Afify	Algoritma SVM dengan *Kernel Linear* menjadi algoritma terbaik yang berhasil memperoleh akurasi sebesar 100%. Sedangkan SVM dengan *Kernel RBF* memperoleh akurasi 99.6%.	Dari penelitian ini menunjukkan bahwa algoritma SVM dapat diimplementasikan dengan menggunakan *Kernel* seperti *Kernel Linear* dan *RBF*.
5	2020 *Elsevier Ltd. All rights reserved,*	*"Coronavirus disease (COVID-19) detection in*	Hasil akurasi SVM menunjukkan bahwa dengan menggunakan	Dari penelitian ini menunjukkan bahwa menggunakan

	Vol -, 2021 [11]	*Chest X-Ray images using majority voting-based classifier ensemble";* Tej Bahadur Chandra, Kesari Verma, Bikesh Kumar Singh, Deepek Jain, Satyabhuwan Singh Netam	*Kernel RBF* diperoleh nilai akurasi sebesar 95.349% dan dengan *Kernel Linear* sebesar 96.124%.	*Kernel* seperti *Kernel Linear* dan *RBF* dapat memberikan hasil yang lebih baik lagi pada model SVM yang dibangun.
6	MDPI, Vol -, 2021 [12]	*"COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA";* Charlyn N. V., Julio J. E. M., Xavier A. I., Jyh-Horng Jeng, Jer-Guang Hsieh	Diperoleh algoritma terbaik yaitu SVM dengan *hyperparameter* optimized dengan nilai *accuracy* sebesar 98,81% dan mean absolute error 0,012.	Penelitian ini menunjukkan algoritma *machine learning* SVM dapat dilakukan optimasi dengan *hyperparameter* untuk memperoleh akurasi tertinggi .
7	2021 The 4th International	"*Support Vector Machine*	Hasil akurasi memodifikasi	Penelitian menunjukkan

	Conference on *Machine learning* and Machine Intelligence, Vol -, 2021 [13]	*Modelling for COVID-19 Prediction based on Symptoms using R Programming Language*"; Charlyn N. Villavicencio, Jyh-Horng, Jeng, Jyh-Horng, Jeng	algoritma SVM dengan *hyperparameter* ini sangat baik yaitu sebesar 98,02%.	bahwa memodifikasi SVM dengan *hyperparameter* juga mampu meningkatkan nilai akurasi penelitian.

Dari tabel 2.3 ini dapat dilihat bahwa terdapat 7 penelitian yang menjadi acuan. Ketujuh penelitian ini tentunya memiliki hubungan atau keterkaitan dengan penelitian yang dilakukan. Pada penelitian [7-9] akan menjadi acuan untuk memilih pemodelan menggunakan algoritma *Support Vector Machine* ini. Lalu pada penelitian [10-11] ini menjadi acuan untuk penggunaan *Kernel Linear* dan *RBF* untuk algoritma *Support Vector Machine*. Sedangkan pada penelitian [12-13] akan menjadi acuan untuk memodifikasi modelnya dengan menggunakan *hyperparameter*.

Namun tentunya terdapat perbedaan dari penelitian terdahulu dengan penelitian yang dilakukan sekarang. Apabila dibandingkan dengan penelitian [7-9], penelitian ini akan menggunakan dataset yang sama dengan algoritma yang sama yaitu SVM. Namun pada ketiga penelitian terdahulu digunakan algoritma SVM biasa (*Simple SVM)*. Sedangkan pada penelitian ini akan dilakukan kombinasi algoritma SVM dengan menggunakan *Kernel Linear, Kernel RBF*, dan *Hyperparameter* dengan menggunakan referensi dari penelitian [10-13].