BAB III
METODOLOGI PENELITIAN

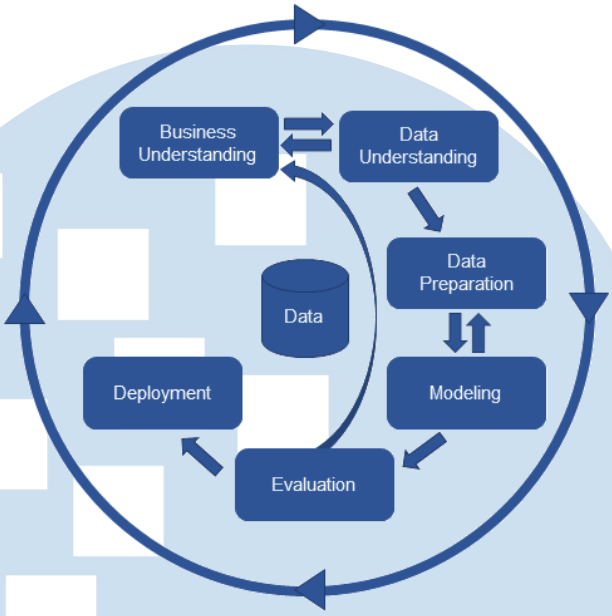
3.1 Gambaran Umum Objek Penelitian

Objek penelitian yang dilakukan dalam penelitian ini berfokus pada analisis prediksi penyakit kanker payudara. Kanker payudara merupakan penyakit yang tumbuh di sel *epitel*, di *duktus*, dan di *lobulus* [1]. Penyakit kanker payudara dapat diketahui dengan mengevaluasi sel kanker dalam tubuh melalui proses yang dinamakan biopsi [17]. Proses deteksi adanya kanker payudara ini dilakukan untuk meminimalisir penyakit kanker payudara agar tidak bertambah parah.

Oleh karena itu penelitian akan berfokus pada data kanker payudara terutama yang diperoleh dari *UCI Machine Learning* dengan judul *Breast Cancer Wisconsin.* Pada dataset yang digunakan terdiri dari 569 data dengan terdiri dari 32 *attribute*. Penelitian ini akan menggunakan keseluruhan datanya dengan memiliki label "*malignant"* dan "*benign*" untuk melakukan deteksi terhadap penyakit kanker payudara ini.

3.2 Metode Penelitian

Pada penelitian ini akan menerapkan atau mengimplementasikan *Cross Industry Standard Process for Data mining* atau biasa disingkat dengan CRISP-DM [19]. CRISP-DM sendiri sudah diakui sebagian besar penulis sebagai standar *de-facto* untuk menerapkan pemodelan dalam membuat *project data mining.* CRISP-DM juga sering digunakan sebagai *framework* dalam meneliti kasus-kasus seperti prediksi atau diagnosis kanker menggunakan pemodelan [21]. Terdapat beberapa penelitian terdahulu yang juga menggunakan *framework* penelitian CRISP-DM [31, 32]. Berikut gambar 3.1 di bawah ini merepresentasikan beberapa tahapan atau proses dari kerangka CRISP-DM yang akan diterapkan pada penelitian.



Gambar 3.1 Tahapan CRISP-DM [20]

3.2.1 *Business Understanding*

Pada tahap *Business Understanding* akan dilakukan *exploration* untuk memahami bagaimana kebutuhan lebih lanjut yang diperlukan yaitu dengan membahas mengenai *objective of problems.* Dengan memahami *objective of problems* nantinya dapat dibuat solusi untuk diterapkan.

Objective of problems pada penelitian ini yaitu pemberitaan mengenai penyakit kanker payudara yang memiliki angka kematian wanita tertinggi terlebih di negara USA [3]. Kematian penyakit kanker payudara ini dikarenakan terlambatnya pendeteksian kanker sehingga penderita terlambat mendapatkan pengobatan dan kanker sudah memasuki stadium lanjut. Upaya untuk deteksi dini penting dilakukan agar dapat cepat ditangai dan memperbesar peluang kesembuhan.

3.2.2 *Data Understanding*

Pada tahapan *Data Understanding* merupakan tahapan yang paling penting untuk memahami dan menganalisis data-data yang akan digunakan. Data yang akan digunakan diperoleh dari sumber *website UCI Machine learning Repository* dengan datasetnya yang berjudul

Breast Cancer Wisconsin. Dataset berjudul *Breast Cancer Wisconsin* ini dibuat oleh *General Surgery Department University of Wisconsin* yaitu Dr. William H. Wolberg beserta *Computer Sciences Department University of Wisconsin* yaitu W. Nick Street dan Olvi L. Mangasarian. Lalu dataset ini didonasikan oleh Nick Street pada 01 November 1995 [37].

Datasetnya memiliki jumlah 32 *attribute* serta 569 data. Keseluruhan data ini dikumpulkan dengan rincian datanya sebanyak 357 untuk *benign* dan 212 untuk *malignant* [37]. Dataset ini juga sangat sesuai dan cocok digunakan untuk penelitian ini karena dataset ini memuat *attribute* yang digunakan untuk mendeteksi penyakit kanker payudara.

3.2.3 *Data Preparation*

Tahapan *Data Preparation* dilakukan untuk mempersiapkan data sehingga dapat memperoleh data yang lebih baik. Dalam tahapan ini akan dilakukan proses *data pre-processing* dan *data split.*

1. *Data pre-processing*

   Pada tahapan *data-preprocessing* akan dilakukan proses *filtering* atau *cleaning* datasetnya. Prosesnya dilakukan dengan mencari data yang memiliki *missing values* atau bernilai NULL untuk segera diperbaiki dengan tidak menggunakan atau menghilangkan baris yang mengandung *missing value.* Pada tahapan ini juga dilakukan pengurangan untuk *attribute* yang tidak digunakan. Dilakukan juga proses *encoding* menggunakan *label encoding* untuk mengubah datasetnya dari categorical ke nilai *integer* serta dilakukan proses normalisasi datanya untuk dipersiapkan ke pemrosesan selanjutnya. Normalisasi data ini dilakukan dengan tujuan untuk mempersiapkan datanya sehingga datanya tidak terlalu besar atau kecil bahkan memiliki range yang sama sehingga proses analisisnya lebih seimbang [38].

2. *Data split*

Pada tahapan *data split* akan dilakukan proses *split* ke dalam data *training* dan data *testing*. Data *training* berperan dalam membangun model sedangkan data *testing* berperan dalam mengukur atau mengevaluasi performa dari model yang digunakan. Proses ini penting dilakukan sebelum mengaplikasikan pemodelan pada data. Pada penelitian ini di *split* datanya dengan perbandingan data *training* sebesar 70% dan data *testing* sebesar 30% mengikuti penelitian terdahulu [13] yang menjadi referensi pada penelitian kali ini.

3.2.4 *Modelling*

Pada tahapan *modelling* akan dilakukan proses pemodelan yang umumnya menerapkan penggunaan metode dan algoritma tertentu untuk membangun pemodelan sebagai solusi untuk memenuhi *Business Understanding.* Proses modelling ini akan menggunakan *tools Jupyter Notebook* dengan menggunakan bahasa pemrograman *Python*. Pemrograman *Python* dipilih karena sudah banyak penelitian terdahulu [7-11] yang menggunakan *Python* dan terbukti Bahasa Pemrograman *Python* lebih efisien.

Pemodelan akan dilakukan dengan menggunakan algoritma *Support Vector Machine* (SVM) dengan *Kernel Linear, Kernel RBF,* dan *hyperparameter* dengan perbandingan fungsinya seperti pada tabel berikut.

Table 3.1 Perbandingan Fungsi Kernel Linear, RBF, Hyperparameter

Pembanding	Fungsi Pembanding
Kernel Linear	*kernel* yang paling sederhana yang berfungsi menganalisis data

	yang terklasifikasi secara *linear* [25]
Kernel RBF	*kernel* yang digunakan dalam suatu data yang biasanya tidak terpisah atau tidak terklasifikasi secara *linear* (*non-linear*) [26]
Hyperparameter	berperan dalam menemukan parameter terbaik dalam pengklasifikasian *machine learning* sehingga memperoleh akurasi terbaik [27]

Pertama, pemodelan dilakukan dengan membuat model SVM dengan menerapkan dua macam *kernel* yaitu *Kernel Linear* dan *Kernel RBF*. Setelah diterapkan *kernel*nya, model SVM nya di *compile* untuk dilakukan *evaluation model* dan dibuat *confusion matrix* nya. Kedua, dilakukan pemodelan kembali dengan algoritma SVM namun menggunakan teknik *hyperparameter*. Setelah selesai dibuat pemodelan, dilakukan proses *evaluation* untuk melihat hasilnya. Kemudian semua tahapan pemodelan ini dibandingkan hasilnya agar dapat ditarik kesimpulan dari metode yang digunakan.

Pemodelan SVM dipilih karena dari penelitian sebelumnya [7-9] algoritma ini berhasil mencapai akurasi tertinggi. Apabila dijabarkan kembali hasil penelitian [7], penelitian ini menggunakan berbagai macam model algoritma diantaranya *Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), k-Nearest Neighbor (KNN),* dan *Decision Tree (DT)*. Saat hasil akhir berupa akurasi dibandingkan dengan algoritma lain misalnya dengan *Random Forest (RF), Logistic Regression (LR), k-Nearest Neighbor (KNN),* dan *Decision Tree (DT),* algoritma *Support Vector Machine (SVM)* lebih

unggul. Karena penelitian ini lebih berfokus mendapatkan akurasi yang sebaik mungkin, maka SVM terpilih menjadi algoritma yang akan digunakan. Sedangkan modifikasi dilakukan dengan *hyperparameter* karena belum ditemukan penelitian dengan dimodifikasi dengan algoritma SVM untuk diterapkan pada data *breast cancer* atau kanker payudara ini. Modifikasi dengan *hyperparameter* juga telah terbukti dapat meningkatkan akurasi apabila digunkan dengan pemodelan lain seperti yang telah dilakukan penelitian [10-11].

3.2.5 *Evaluation*

Dalam tahapan *evaluation* akan dilakukan proses evaluasi performa dari pemodelan yang telah diterapkan. Proses evaluasi ini berperan untuk mengetahui bagaimana hasil prediksi dengan pemodelan yang telah dibangun untuk menyelesaikan *Business Understanding*. Tahapan *evaluation* akan dilakukan dengan menampilkan nilai akurasi, *classification report, confusion matrix,* serta hasil dari pengujian model *data mining* dengan *data testing* yang sudah dibagi sebelumnya mengikuti penelitian terdahulu [7-11]. Kemudian hasil akurasinya akan dilakukan analisis lebih lanjut dan akan dibandingkan dengan akurasi penelitian terdahulu [7-9].

3.2.6 *Deployment*

Pada penelitian ini, tidak dilakukan hingga ke tahapan *deployment*. Ini dikarenakan penelitiannya hanya sebatas untuk keperluan studi saja dan tidak diimplementasikan pada suatu divisi perusahaan.

3.3 Variabel Penelitian

Penelitian ini menggunakan dataset dari *UCI Machine learning Repository* berjudul *Breast Cancer Wisconsin* [37] sebab *UCI Machine learning Repository* sudah dikutip lebih dari 1000 kali di berbagai jurnal baik jurnal nasional maupun internasional [39]. Datasetnya berisikan 32 *attribute* yang akan terbagi menjadi variabel independen dan dependen untuk mendeteksi kanker payudara.

3.3.1 Variabel Independen

Variabel independen merupakan variabel yang dapat memberikan dampak atau memberikan pengaruh pada variabel lainnya [40]. Pada penelitian dengan dataset *Breast Cancer Wisconsin*, terdapat variabel independen yang mempengaruhi yang terdiri dari *attribute id, radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean, radius_se, texture_se, perimeter_se, area_se, smoothness_se, compactness_se, concavity_se, concave points_se, symmetry_se, fractal_dimension_se, radius_worst, texture_worst, perimeter_worst, area_worst, smoothness_worst, compactness_worst, concavity_worst, concave points_worst, symmetry_worst, fractal_dimension_worst* [7-9].

3.3.2 Variabel Dependen

Variabel dependen merupakan variabel yang terkena dampak atau dipengaruhi dari variabel independen [40]. Pada penelitian dengan dataset *Breast Cancer Wisconsin*, variabel dependennya terdapat pada *attribute diagnosis* [7-9].

3.4 Teknik Pengumpulan Data

Teknik pengumpulan data merupakan suatu metode yang dibutuhkan untuk mengumpulkan data untuk memenuhi suatu permasalahan pada penelitian [41]. Diketahui bahwa terdapat teknik pengumpulan data primer, data sekunder, dan data tersier. Data primer yaitu data yang dikumpulkan langsung dari sumbernya untuk pertama kalinya, data sekunder adalah data yang diperoleh dari studi literatur berbagai sumber, sedangkan data tersier adalah data penunjang yang biasanya diperoleh melalui kamus, ensiklopedia, ataupun dari sumber lain yang masih memiliki keterkaitan dengan masalah yang diteliti [42].

Pada penelitian ini pengumpulan data dilakukan secara tersier dengan diperoleh dari *website UCI Machine learning Repository* yang berjudul *Breast Cancer Wisconsin* [43]. Data *Breast Cancer Wisconsin* ini Data yang

dikumpulkan berupa data kuantitatif karena datanya disajikan dalam bentuk angka dan dapat di ukur.

3.5 Teknik Pengambilan Sampel

Pengambilan sampel data dilakukan dengan mengambil data secara tersier dari *UCI Machine learning* [43]. Sampel data keseluruhannya diambil dari *UCI Machine learning* karena sudah banyak penelitian yang mempercayakan sumber tersebut. Selain itu sudah terjamin kredibilitas datanya sebab sudah banyak penelitian terdahulu yang menggunakan data tersebut [7-9].

3.6 Teknik Analisis Data

Pada analisis datanya akan menerapkan pencarian nilai akurasi. Analisis menggunakan nilai akurasi dilakukan sebagaimana yang telah dilakukan oleh penelitian terdahulu [7-11]. Nantinya nilai akurasi akan digunakan untuk mengevaluasi pemodelan yang telah dibuat. Penelitian ini dapat dikatakan baik apabila memperoleh nilai akurasi diatas 80% [43] atau memperoleh nilai akurasi yang lebih tinggi dari penelitian sebelumnya.