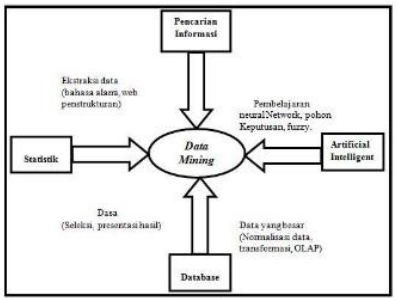
BAB II

LANDASAN TEORI

2.1 Data Mining

Menurut Roiger, R. J. (2017:20) dalam buku yang berjudul *"Data Mining A Tutorial-Based Primer"* mendefinisikan bahwa "Data mining sebagai proses menemukan struktur yang menarik dalam data. Struktur dapat mengambil banyak bentuk, termasuk seperangkat aturan, grafik atau jaringan, pohon, satu atau beberapa persamaan, dan lainnya.".

Menurut Sianturi, F. A. (2017) Data mining adalah proses yang mempekerjakan satu atau lebih teknik pembelajaran komputer *(mechine learning)* untuk menganalisis dan mengekstraksi pengetahuan *(knowledge)* secara otomatis. Defenisi lain diantaranya adalah pembelajaran berbasis induksi *(induction-based learning)* adalah proses pembentukan defenisi-defenisi konsep umum yang dilakukan dengan cara mengobservasi contoh-contoh spesifik dari konsep-konsep yang akan dipelajari. *Knowledge Discovery in Databases* (KDD) adalah penerapan metode saintifik pada data mining. Dalam konteks ini data mining merupakan satu langkah dari proses KDD.



Gambar 2.1 Bidang Ilmu Data Mining

Sumber : Sianturi, F. A., 2017

2.2 Prediksi

Menurut Wanto, A., & Windarto, A. P. (2017) “Prediksi adalah usaha menduga atau memperkirakan sesuatu yang akan terjadi di waktu mendatang dengan memanfaatkan berbagai informasi yang relevan pada waktu-waktu sebelumnya (historis) melalui suatu metode ilmiah.”.

2.3 Tenaga Kerja

Siswanto dalam Sumolang dkk. (2019) mendiskusikan bahwa “Tenaga Kerja merupakan istilah yang identik dengan istilah personalia, di dalamnya meliputi buruh. Buruh yang dimaksud adalah mereka yang bekerja pada usaha perorangan dan diberikan imbalan kerja secara harian maupun borongan sesuai dengan kesepakatan kedua belah pihak, biasanya imbalan kerja tersebut diberikan secara harian.”.

2.4 CRISP-DM

Menurut Bernardus Ari Kuncoro (2020) pada buku yang berjudul “Pengenalan Prinsip *Data Science* untuk Pemula” mendefinisikan bahwa “CRISP-DM merupakan singkatan dari *Cross-Industry Standard Process for Data Mining*. Sebuah metodologi yang menerapkan pendekatan terstruktur untuk perencanaan proyek data mining yang sangat ampuh dan sudah teruji dengan baik. Metode ini sangat umum digunakan karena sangat praktis, fleksibel, dan aplikatif untuk memecahkan isu bisnis yang sulit sekalipun. Metode ini merupakan metode andalan yang dapat dijalankan di hampir semua persoalan bisnis data mining.”.

Gambar 2.2 Proses CRISP-DM
Sumber : Bernardus Ari Kuncoro, 2020

Metode CRISP-DM terdiri dari enam (6) tahapan yaitu :

1. Pemahaman Bisnis *(Business Understanding)*

   Tahapan ini menfokuskan pada pemahaman tujuan projek dan kebutuhan-kebutuhan yang diinginkan bisnis, kemudian merubahnya pengetahuan ini untuk mendefinisikan data mining dan rencana yang ingin dilakukan untuk mencapai tujuan bisnis.

2. Pemahaman Data *(Data Understanding)*

   Tahapan memahami data dimulai dengan mengumpulkan data awal dan dilanjutkan dengan kegiatan-kegiatan untuk mendapatkan data yang lazim serta identifikasi data yang berkualitas, pemahaman data sangat diperlukan untuk mendeteksi bagian yang menarik dari data sehingga dapat membangun hipotesa terhadap informasi yang tersembunyi.

3. Persiapan Data *(Data Preparation)*

   Fase *data preparation* ini memerlukan pemikiran matang dan upaya tinggi untuk memperbaiki masalah dalam data dan dibuat *variable derived* serta memastikan apakah data sudah tepat untuk

algoritma yang digunakan. Tahap ini sering mengalami peninjauan ulang ketika menemukan kendala pada pembangunan model, sehingga dilaksanakan iterasi hingga menemukan hal yang sesuai dengan data yang dimaksud.

4. Pemodelan *(Modeling)*

   Pada tahap ini, dilakukan metode statistika dan *machine learning* untuk menentukan teknik, alat bantu serta algoritma data mining yang akan diterapkan. Kemudian langkah selanjutnya adalah menerapkan teknik dan algoritma tersebut pada data dengan alat bantu. Yang perlu digaris bawahi disini, beberapa teknik memungkinkan untuk digunakan pada data mining yang memiliki permasalahan yang sama. Jika diperlukan penyesuaian data terhadap metode data mining, kita dapat kembali ke tahapan *data preparation.*

5. Evaluasi *(Evaluation)*

   Tahapan *evaluation* ini merupakan tahap evaluasi dengan melaksanakan interpretasi terhadap *output* dari data mining yang dihasilkan dalam tahapan sebelumnya. Evaluasi disini bertujuan agar model yang sudah ditentukan dapat sesuai dengan tujuan yang ingin dipenuhi pada fase pertama.

6. Penyebaran *(Deployment)*

   Tahap *deployment* atau rencana penggunaan model merupakan fase yang penting dalam proses CRISP-DM. Perencanaan untuk tahap *deployment* dimulai sejak proses *Business*

Understanding dilakukan. Fase *deployment* ini tidak hanya menghasilkan suatu model, tapi juga mengonversi skor putusan serta menggabungkan keputusan dalam sistem operasional.

Pada akhirnya, rencana sistem *deployment* mengakui bahwa tidak ada model yang statis. Model tersebut dibangun dari data yang diwakili data pada waktu tertentu, sehingga perubahan waktu dapat menyebabkan berubahnya karakteristik data. Modelpun harus dipantau dan mungkin diganti dengan model yang sudah diperbaiki.

2.5 *Website*

2.5.1 *CSS (Cascading Style Sheets)*

Richard Blum (2018, h. 11) mendefinisikan "Seperangkat aturan yang menentukan bagaimana browser harus menerapkan *style* ke dokumen HTML.".

Rohi Abdulloh (2018, h. 3) dalam bukunya mengatakan "Sebagai pembentuk desain website dengan mengatur setiap elemen website sesuai *layout* yang diinginkan.".

2.5.2 *Javascript*

Vivian Siahaan & Rismon Hasiholan (2018, hal. 1) dalam buku "*JavaScript* dari A sampai Z" mendiskusikan tentang *JavaScript*. Dan mereka mengatakan bahwa "*JavaScript* merupakan bahasa skript populer yang dipakai untuk menciptakan halaman web yang dapat berinteraksi dengan pengguna dan dapat merespon *event* yang terjadi pada halaman. *JavaScript* merupakan perangkat yang menyatukan halaman-halaman *web*.".

2.5.3 PHP *(Hypertext Preprocessor)*

Rohi Abdulloh (2018, h. 3) mengatakan bahwa "PHP berperan sebagai proses data pada sisi *server*, sesuai yang diminta oleh *client* menjadi informasi yang siap ditampilkan, juga sebagai penghubung aplikasi *web* dengan *database*.".

2.5.4 *Bootstrap*

Zaenal A Rozi & SmitDev Community (2015, h. 1) mendiskusikan topik *bootstrap*. Dan mereka mengatakan bahwa "*Bootstrap* adalah paket aplikasi siap pakai untuk membuat *front-end* sebuah *website*. Bisa dikatakan, *bootstrap* adalah template desain *web* dengan fitur plus. *Bootstrap* diciptakan untuk mempermudah proses desain *web* bagi berbagai tingkat pengguna, mulai dari level pemula hingga yang sudah berpengalaman.".

2.5.5 HTML *(Hypertext Markup Language)*

Rohi Abdulloh (2018, h. 7) mengatakan bahwa "Bahasa standar *web* yang dikelola penggunaannya oleh W3C *(World Wide Web Consortium)* berupa tag-tag yang menyusun setiap elemen dari website. HTML. Berperan sebagai penyusun struktur halaman website yang menempatkan setiap elemen website sesuai layout yang diinginkan.".

2.6 *Visual Studio Code (VS Code)*

Ummy Gusti Salamah (2021, h. 1) mendefinisikan bahwa "*Visual Studio Code (VS Code)* ini adalah sebuah teks editor ringan dan handal yang dibuat oleh Microsoft untuk system operasi multiplatform, artinya tersedia

juga untuk Linux, Mac, dan Windows. Teks editor VS *Code* juga bersifat open source, yang mana kode sumbernya dapat kalian lihat dan kalian dapat berkontribusi untuk pengembangannya. Hal ini juga yang membuat VS *Code* menjadi favorit para pengembang aplikasi, karena para pengembang aplikasi bisa ikut serta dalam proses pengembangan VS *Code* ke depannya.

2.7 Database

2.7.1 MySQL

Yoga Ananda Putra, Sumijan, & Mardison. (2019) mendiskusikan bahwa "MySQL adalah sebuah implementasi dari sistem manajemen basis data relasional (RDBMS) yang didistribusikan secara gratis dibawah lisensi GPL *(General Public License)*. Setiap pengguna dapat secara bebas menggunakan MySQL, namun dengan batasan perangkat lunak tersebut tidak boleh dijadikan produk turunan yang bersifat komersial.

MySQL sebenarnya merupakan turunan salah satu konsep utama dalam basisdata yang telah ada sebelumnya; SQL *(Structured Query Language)*. SQL adalah sebuah konsep pengoperasian basis data, terutama untuk pemilihan atau seleksi dan pemasukan data, yang memungkinkan pengoperasian data dikerjakan dengan mudah secara otomatis.

2.8 XAMPP

Darman Umagapi & Arisandy Ambarita (2018) mendiskusikan bahwa "XAMPP adalah perangkat lunak *(free software)* bebas, yang

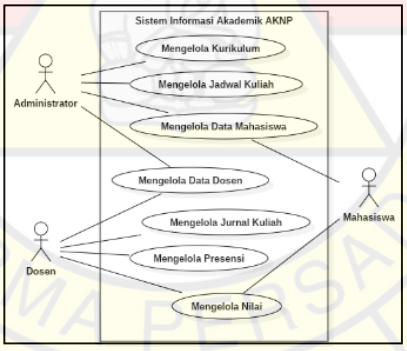
mendukung untuk banyak sistem operasi, yang merupakan kompilasi dari beberapa program.”.

2.9 UML

Menurut Sulianta (2017) dalam buku “Teknik Perancangan Arsitektur Sistem Informasi” mengatakan bahwa “*Unified Modeling Language* (UML) merupakan kumpulan diagram-diagram yang sudah memiliki standar untuk membangun perangkat lunak berbasis objek.”.

Menurut Sukamto dan Shalahuddin (2018:133), mendiskusikan bahwa, “UML merupakan sebuah standar bahasa yang digunakan untuk menganalisis dan merancang serta menggambarkan arsitektur program dalam pemrograman *object oriented.*”.

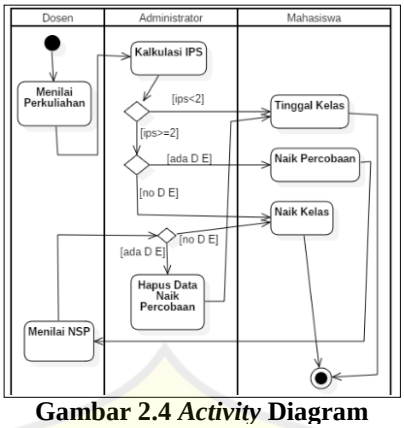
2.9.1 *Use Case* Diagram



Gambar 2.3 *Use Case* Diagram
Sumber : Fu'adi, A., & Prianggono, A., 2022

Yuni Sugiarti (2018, h. 108) mengatakan bahwa *“Use case* diagram merupakan pemodelan untuk menggambarkan *behavior* dan mendeskripsikan sebuah interaksi antara satu atau lebih aktor dengan sistem yang akan dibuat.”.

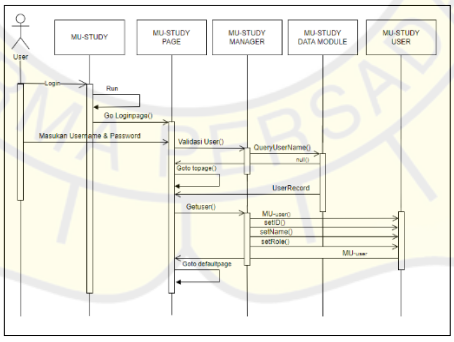
2.9.2 *Activity* Diagram



Gambar 2.4 *Activity* Diagram
Sumber : Fu'adi, A., & Prianggono, A., 2022

Menurut Yuni Sugiarti (2018, h. 133) *"Activity* diagram menggambarkan *workflow* (aliran kerja) atau aktivitas dari sebuah sistem atau proses bisnis. Hal yang perlu diperhatikan di sini adalah bahwa diagram aktivitas menggambarkan kegiatan sistem bukan apa yang dilakukan aktor, jadi aktivitas yang dapat dilakukan oleh sistem.".

2.9.3 *Sequence* Diagram

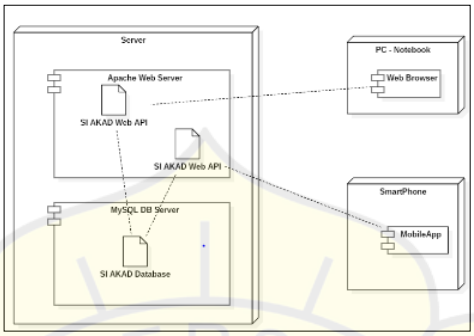


Gambar 2.5 *Sequence* Diagram
Sumber : Ricky Rohmanto & Topan Setiawan, 2022

Yuni Sugiarti (2018, h. 130) mengatakan bahwa "Diagram sekuens (*sequence)* menggambarkan behavior objek pada *Use case* dengan mendeskripsikan waktu hidup dan message yang dikirimkan dan diterima antar objek. Banyaknya diagram sekuens yang harus

digambar adalah sebanyak pendefinisian *Use case* yang memiliki proses sendiri atau yang penting semua *Use case* telah didefinisikan interaksi jalannya pesan sudah dicakup pada diagram sekuens.”.

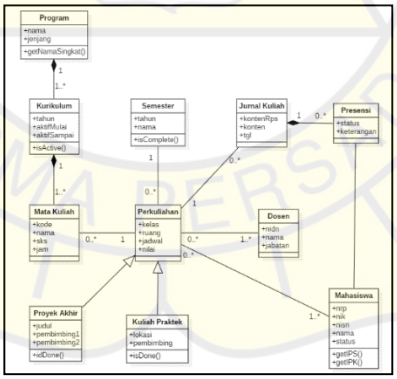
2.9.4 *Deployment* Diagram



Gambar 2.6 *Deployment* Diagram
Sumber : Fu'adi, A., & Prianggono, A., 2022

Yurindra (2017, h. 32) mengatakan bahwa “*Deployment* diagram merupakan gambaran proses-proses berbeda pada suatu sistem yang berjalan dan bagaimana relasi di dalamnya.”.

2.9.5 *Class* Diagram



Gambar 2.7 *Class* Diagram
Sumber : Fu'adi, A., & Prianggono, A., 2022

Menurut (Sukamto & Shalahuddin, 2018) mengemukakan bahwa, “*Class* diagram menggambarkan struktur sistem dari segi pendefinisian kelas-kelas yang akan dibuat untuk membangun *system*”, sedangkan menurut (Tohari, 2014), “*Class* diagram mendeskripsikan

jenis-jenis objek dalam sistem dan berbagai macam hubungan statis yang terdapat diantara mereka juga menunjukkan properti dan operasi sebuah class dan batasan-batasan yang terdapat dalam hubungan-hubungan objek".

2.10 *Decision Tree*

Decision Tree merupakan algoritma yang umum digunakan untuk pengambilan keputusan dengan membentuk cabang-cabang dari setiap keputusan. (Sartika et al.2017)

Menurut Anief Rufianto, M. Rochcham dan Abdul Rohman dalam buku yang berjudul "Penerapan Algoritma C4.5 Untuk Prediksi Kepuasan Mahasiswa Tahun 2020" mendiskusikan bahwa "Decision tree atau pohon keputusan adalah metode atau algoritma klasifikasi data mining dengan membentuk pola pohon keputusan yang digunakan untuk mendapatkan jawaban dari masalah yang dimasukan. Dengan pohon keputusan, dapat dengan mudah mengidentifikasikan hubungan antar faktor-faktor yang mempengaruhi masalah dan mencari solusi yang baik dengan memperhitungkan faktor-faktor tersebut.".

2.10.1 Algoritma C4.5

Menurut Damanik, S. F., Wanto, A., & Gunawan, I. (2022), Algoritma C4.5 adalah salah satu metode klasifikasi dari data mining yang digunakan untuk mengkonstruksikan pohon keputusan (*Decision Tree*). Algoritma C4.5 adalah program yang memberi kontribusi satu set data berlabel dan menghasilkan pohon keputusan sebagai keluaran. Pohon keputusan tindak lanjut ini kemudian diverifikasi terhadap data

uji berlabel yang tidak terlihat untuk menghitung generalisasinya. C4.5 adalah program yang digunakan untuk menghasilkan peraturan taksonomi dengan menggunakan pohon keputusan dari sekumpulan data yang diberikan.

Algoritma C4.5 merupakan perpanjangan dari algoritma ID3 dasar dan dirancang oleh Quinlan. C4.5 adalah salah satu algoritma pembelajaran yang banyak digunakan. Algoritma C4.5 membangun pohon keputusan dari serangkaian data pelatihan yang mirip dengan Algoritma ID3, dengan menggunakan konsep entropi informasi. C4.5 juga dikenal sebagai klasifikasi statistik.

Algoritma C4.5 merupakan metode yang menjadi pilihan pertama dan sering digunakan dalam pengembangan *Data Mining* karena kecepatan dalam pengklasifikasian pohon keputusan disamping dapat mengkonstruksi pengklasifikasian dengan aturan-aturan yang lain. Algoritma ini mempunyai *input* berupa *training* samples dan samples. *Training* samples berupa data contoh yang akan digunakan untuk membangun sebuah tree yang telah diuji kebenarannya. Sedangkan samples merupakan *field-field* data yang nantinya akan digunakan sebagai parameter dalam melakukan klasifikasi data.

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut.

a. Pilih atribut sebagai akar.
b. Buat cabang untuk tiap-tiap atribut.

c. Bagi kasus dalam cabang.

d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kasus yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai *gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *gain* digunakan rumus seperti tertera pada persamaan 1 berikut.

$$Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i) \ldots (1)$$

Keterangan :

S : Himpunan kasus

A : Atribut

N : Jumlah partisi atribut A

|Si| : Jumlah kasus pada partisi ke-i

S : Jumlah kasus dalam S

Entropi(S) merupakan jumlah bit yang diperkirakan dibutuhkan untuk dapat mengekstrak suatu kelas dari sejumlah data acak pada ruang sampel S. Entropy dapat dikatakan sebagai kebutuhan bit untuk menyatakan suatu kelas. Pada persamaan (2) merupakan rumus entropi :

$$Entropy(S) = \sum_{i=1}^{n} - p_i * log_2 p_i \ldots (2)$$

Dimana :

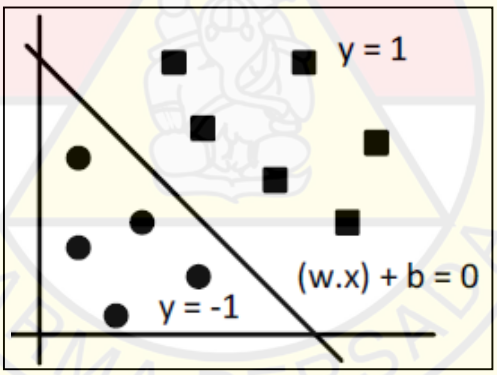
S : Ruang (data) sampel yang digunakan untuk pelatihan.

n : Jumlah partisi S

pi : Proporsi dari Si terhadap S

2.11 *Support Vector Machine* (SVM)

Pada SVM dapat mengklasifikasikan data linier dan non linier. Data *input* merupakan nilai variabel-variabel prediktor sedangkan *output* merupakan variabel target dimana saling bergantung. Dalam tujuan SVM adalah untuk menemukan fungsi klasifikasi terbaik untuk membedakan antara anggota dari dua kelas dalam data pelatihan. Metrik untuk konsep fungsi klasifikasi "terbaik" dapat diwujudkan secara geometris. Untuk *dataset* terpisah secara linear, fungsi klasifikasi linier berhubungan dengan *hyperplane* pemisah f (x) yang melewati tengah dua kelas, memisahkan keduanya (Neelamegam & Ramaraj, 2013).



Gambar 2.8 Bidang Pemisah Linier

2.12 *Confusion Matrix*

Menurut Nawangsih, I., & Fauziah, S. (2021) mendiskusikan bahwa "*Confusion Matrix* adalah tool yang digunakan untuk evaluasi model klasifikasi untuk memperkirakan objek yang benar atau salah. Sebuah matrix dari prediksi yang akan dibandingkan dengan kelas yang asli dari *input*an atau dengan kata lain berisi informasi nilai aktual dan prediksi pada klasifikasi.".

Menurut Harani, N. H., & Hasanah, M. (2020) dalam buku yang berjudul "Deteksi Objek dan Pengenalan Plat Nomor Kendaraan Indonesia Berbasis Pyhton" mendiskusikan bahwa "*Confusion Matrix* adalah salah satu metode yang dapat digunakan untuk mengukur kinerja suatu metode klasifikasi dan pada dasarnya Confusion Matrix mengandung informasi yang dapat membandingkan hasil klasifikasi yang seharusnya.".

Kelas	Ter klasifikasi Positif	Ter klasifikasi Negatif
Positif	TP (True Positive)	FN (False Negative)
Negatif	FP (False Positive)	TN (True Negative)

Gambar 2.9 *Confusion Matrix*
Sumber : Harani, N. H., & Hasanah, M., 2020

Berdasarkan nilai *False Positive* (FP), *False Negative* (FN), *True Negative* (TN), dan *True Positive* (TP) dapat diperoleh nilai akurasi, presisi dan *recall*. Nilai dari akurasi menggambarkan seberapa akurat/cermat suatu sistem dapat mengklasifikasikan data secara benar. Dengan kata lain, nilai akurasi merupakan perbandingan antara data yang terklasifikasi benar dengan keseluruhan data.

Pada tahap ini akan dijelaskan bagaimana nilai *Confusion Matrix* didapat berdasarkan penggunaan rumus-rumus yang ada dibawah ini, sebagai berikut.

Accuracy = (TP + TN) / (TP + FP + TN + FN)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 Score = (2 * Precision * Recall) / (Precision + Recall)