BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Tabel 2. 1 Tabel Penelitian Sebelumnya

No	Penulis	Judul	Jurnal	Hasil	Kesimpulan
1	Imron Rosydi, Aryo Nugroho, Awalludiyah Ambarwati (2022) [9].	Sistem Monitoring BTS Pada Perusahaan Telekomunikasi Seluler Berbasis Aplikasi *Mobile*	JOINTECS (*Journal of Information Technology and Computer Science*)	Hasil penelitian akan diuji menggunakan 3 karakteristik ISO 25010 yakni *suitability*, *usability*, dan *performance efficiency*. Berdasarkan pengujian penelitian ini telah memenuhi karkteristik ISO 25010 [9].	Sistem monitoring berbasis *mobile* telah memenuhi uji karkteristik ISO 25010, dan dapat membantu dalam menangani BTS yang sedang mengalami kendala.
2	Marta Kuc-Czarnecka (2020) [10]	Covid-19 *and digital deprivation in Poland*	*Oeconomia Copernicana*	Penelitian ini membahas dampak dari pandemi Covid-19 terhadap eksklusi digital di Polandia. Penelitian ini membahas mengenai kekurangan akses digital yang didorong oleh pandemi Covid-19. Penelitian ini menggunakan konsep GIS dan BTS untuk menemukan berbagai daerah yang membutuhkan bantuan akses internet [10].	Pandemi Covid-19 memiliki peran yang siknifikan dalam mengungkapkan serta memperburuk masalah kekurangan digital di Polandia. Pandemi ini memaksa berbagai aktivitas untuk dilakukan secara *online* di mana selain menghambat bagi warga yang tidak memiliki akses internet juga menghambat rakyat yang memiliki akses internet dikarenakan koneksi yang tidak stabil.
3	Jorge E. Preciado-Velasco,	*5G/B5G Service Classification Using*	*Applied Sciences (Switzerland)*	Dalam penelitian ini akan dilakukannya prediksi menggunakan	Penelitian ini memberikan kesimpulan bahawa adanya potensi analisis

	Joan D. Gonzalez-Franco, Caridad E. Anias-Calderon, Juan I. Nieto-Hipolito, Raul Rivera-Rodriguez (2021) [8].	*Supervised Learning*		*machine learning* terhadap jaringan 5G/B5G berdasarkan metrik KPI. Hasil analisis ini akan berbentuk prediksi layanan yang cocok untuk metrik KPI yang digunakan. Penelitian ini mendapatkan Random Forest dan Decision Tree sebagai model optimal dengan akurasi 96.9% tanpa adanya *overfitting*. Penelitian ini juga menemukan algoritma SVM mendapatkan akurasi 100% tetapi setelah uji coba *cross validation* ditemukan perbedaan hasil lebih dari 5% yang menunjukkan adanya *overfitting* [8].	menggunakan *machine learning* dalam melakukan prediksi layanan telekomunikasi 5G/B5G. Dengan menggunakan metrik KPI dan KQI menunjukkan model akan lebih optimal dikarenakan jumlah data yang lebih banyak akan berpengaruh terhadap hasil pelatihan model.
4	Natanael Benediktus, Raymond Sunardi Oetama [11]	Algoritma Klasifikasi Decision Tree C5.0 untuk Memprediksi Performa Akademik Siswa	ULTIMATICS	Penelitian ini membahas implementasi algoritma Decision Tree terhadap performa akademik mahasiswa. Penelitian ini melakukan prediksi terhadap performa akademik mahasiswa berdasarkan beberapa faktor seperti absensi dan diskusi [11].	Penelitian ini menggunakan algoritma Decision Tree untuk memprediksi performa akademik mahasiswa. Berdasarkan hasil *confusion matrix* serta akurasi yang tertulis dapat disimpulkan algoritma Decision Tree cukup efektif untuk memprediksi performa akademik mahasiswa untuk seluruh faktor dengan akurasi 71%

5	Alexander Bryan Wiratman, Wella (2024) [12]	*Personalized Learning Models Using Decision Tree and Random Forest Algorithms in Telecommunication Company*	*International Journal on Informatics Visualization*	Penelitian ini melakukan pembuatan sistem penilaian karyawan menggunakan *dashboard* dan *machine learning*. Dashboard digunakan untuk menunjukkan hasil pelatihan karyawan sedangkan *machine learning* digunakan untuk memberikan rekomendasi pelatihan berdasarkan hasil prediksi. Model yang digunakan adalah Random Forest dengan akurasi 70% dan Decision Tree dengan akurasi 69% [12].	Penelitian ini membahas terkait pembuatan sistem penilaian karyawan untuk salah satu perusahaan telekomunikasi menggunakan *dashboard* visualisasi data dan model *machine learning*. *Dashboard* dibuat menggunakan Tableau yang menampilkan hasil pelatihan karyawan. *Machine learning* digunakan untuk memberikan rekomendasi pelatihan yang dipersonalisasi. Sistem ini dapat memberikan lingkungan belajar yang efektif dan efisien.
6	Abdelrahim Kasem Ahmad, Assef Jafar, Kadan Aljoumaa (2019) [7]	*Customer churn prediction in telecom using machine learning in big data platform*	*Journal of Big Data*	Penelitian ini membahas terkait penggunaan *machine learning* untuk memprediksi *churn* pelanggan dalam perusahaan telekomunikasi. Penelitian ini menggunakan ROC-AUC sebagai pembanding dan menggunakan berbagai algoritma untuk dibandingkan seperti XGBoost dengan *score* 93.3%, GSM 90.89%, Random Forest 78.47%, dan Decision Tree 72.2% [7].	Hasil penelitian ini menggunakan data perusahaan SyriaTel yang menyediakan beberapa data terkait pelanggan perusahaan. Hasil analisis yang dilakukan ditemukan bahwa algoritma XGBoost lebih optimal dalam memprediksi *churn* pelanggan telekomunikasi.
7	Deasy Arisanty, Muhammad	*Spatiotemporal Patterns of Burned Areas*	*Journal of Forestry Research Volume 2021*	Penelitian ini membahas terkait pembuatan sistem menggunakan	Penelitian ini menggunakan metode cluster untuk membantu

	Muhaimin, Dedi Rosadi, Aswin Nur Saputra, Karunia Puji Hastuti, Ismi Rajiani (2021) [13]	*Based on the Geographic Information System for Fire Risk Monitoring*		metode *hotspot* dan *kernel density analysis* untuk mencari daerah rawan kebakaran di Indonesia [13].	mencari daerah kalimantan yang rawan terjadi kebakaran serta intensitas kebakaran yang dapat terjadi menggunakan metode *hot spot analysis* dan *kernel density analysis*. Area yang rawan kebakaran umumnya terjadi di area yang memiliki kepadatan vegetasi serta terjadi di berbagai area lainnya seperti sungai Hulu.
8	Bo Liu (2023) [14]	*Based on intelligent advertising recommendation and abnormal advertising monitoring system in the field of machine learning*	*International Journal of Computer Science and Information Technology*	Penelitian ini membahas terkait pembuatan sistem pemantauan periklanan online dengan *machine learning*. Penelitian ini menggaunakan *machine learning* untuk mempelajari pola data agar dapat memiliki sistem pemantauan yang lebih optimal [14].	Penelitian ini membuat dua buah sistem yakni *Intelligent Advertising Recommendation* yang bertujuan memberikan iklan yang menyesuaikan dengan konsumen dan *Abnormal Advertising Monitoring* yang digunakan untuk mencari mendeteksi pola abnormal dalam data iklan. Sistem ini memungkinkan penunjukkan iklan yang sesuai dengan keinginan dan kebutuhan sebuah konsumen sesuai data yang sudah diberikan.
9	Partha Protim Roy, Md. Shahriar Abdullah, Iqtiar Md.	*Machine learning empowered geographic information systems:*	*World Journal of Advanced Research and Review*	Penelitian ini membahas integrasi antara *machine learning* dan GIS. Penelitian ini membahas terkait	Penelitian ini memberikan kesimpulan bahwa sistem integrasi GIS dan *machine learning*

	Siddique (2024) [15]	*Advancing Spatial analysis and decision making*		penggunaan model prediktif *machine learning* dan menggabungkannya dengan teknologi GIS yang dapat memberikan gambaran sesuai lokasi yang *real* [15].	memiliki potensi yang besar dalam analisis spasial dan *decision making*. *Machine learning* membantu tahap analisis spasial yang kompleks serta menyediakan kemampuan model predikitf. GIS digunakan untuk menampilkan hasil prediksi sehingga dapat memberikan informasi yang lebih luas.
10	Bahzad Taha Jijo , Adnan Mohsin Abdulaze ez (2021) [16]	*Classification Based on Decision Tree Algorithm for Machine Learning*	*Journal of Applied Science and Technology Trends.*	Penelitian ini membahas penggunaan algoritma Decision Tree dalam klasifikasi teks. Penelitian ini akan melakukan komparasi terhadap berbagai algoritma *machine learning* untuk mengevaluasi efisiensi algoritma tersebut. Penelitian ini juga membahas secara detail terkait Decision Tree dan berbagai teknik yang dapat dilakukan untuk meningkatkan efisiensi dan mencegah *overfitting*. Akurasi terbaik yang dicapai oleh Decision Tree adalah 99.93% dengan menggunakan *machine learning repository dataset* [16].	Decision Tree merupakan algoritma *machine learning* yang populer serta efektif untuk melakukan klasifikasi data. Terdapat berbagai metode yang dilakukan untuk meningkatkan efisiensi model seperti menggunakan konsep *pruning* untuk mencegah *overfitting* dan menggeneralisasi data baru.

Tabel 2.1 menunjukkan beberapa penelitian terdahulu yang dijadikan refresni untuk penelitian yang akan dilakukan. Secara keseluruhan penelitian ini akan memiliki perbedaan dengan penelitian sebelumnya khususnya pada tahap pembangunan sistem atau tahap *deployment*. Pada penelitian ini akan melakukan pembangunana sistem berbasis peta yang disertakan juga hasil analisis prediktif dari *machine learning*. Penelitian ini dapat membuktikan serta memberikan hasil adanya potensi dengan menggabungkan kemampuan prediktif dari *machine learning* dengan kemampuan informatif dan visibilitas dari peta.

Terdapat beberapa penelitian terdahulu yang dapat dijadikan acuan untuk penelitian ini. Penelitian *5G/B5G Service Classification Using Supervised Learning* membahas terkait pengguna *supervised machine learning* untuk memprediksi layanan jaringan 5G/B5G berdasarkan metrik KPI dan KQI [8]. Penelitian *Customer churn prediction in telecom using machine learning in big data platform* membahas prediksi *churn* pelanggan untuk perusahaan telekomunikasi SyriaTel [7]. Kedua penelitian tersebut membahas salah satu aspek telekomunikasi selain itu juga menggunakan metode *supervised learning* yakni *classification* untuk membantu prediksi masalah yang sedang dialami. Penelitian ini juga memberikan hasil berdasarkan hasil evaluasi model yang dilakukan dan tidak melanjutkan ke tahap berikutnya seperti penerapan model kedalam sebuah sistem. Berbeda dengan penelitian sebelumnya, pada penelitian ini akan menerapkan model yang paling optimal kedalam sebuah sistem pemantauan berbasis peta untuk pihak perusahaan.

Pada penelitian "*Spatiotemporal Patterns of Burned Areas Based on the Geographic Information System for Fire Risk Monitoring*." Membahas pembuatan sistem berbasis GIS untuk memantau daerah di Indonesia yang rawan terjadi kebakaran. Pada penelitian ini memiliki fokus utama pembuatan sistem GIS berdasarkan data yang sudah disediakan oleh *National Oceanic and Atmospheric Administration* periode 2016-2019 [13]. Penelitian ini menggunakan metode *hotspot* dan *kernel density* yang

berbeda dengan metode yang akan digunakan dalam penelitian *data mining*. Penelitian yang akan dilakukan, akan menggunakan metode *classification machine learning* untuk analisis data, lalu menggunakan hasil model optimal untuk sistem berbasis peta yang akan dirancang.

2.2 Landasan Teori

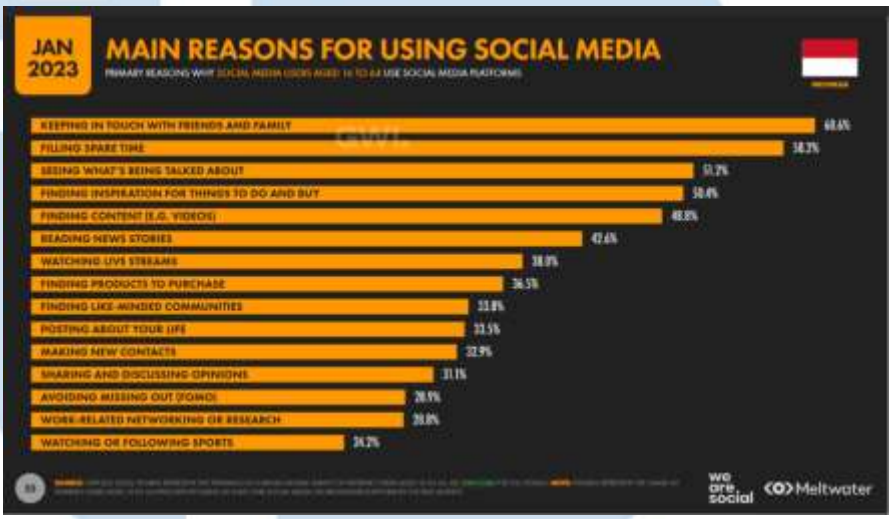
2.2.1 Komunikasi

Sebagai makhluk sosial manusia saling membutuhkan satu sama lain, sehingga manusia tidak akan pernah lepas dari komunikasi [17]. Komunikasi pada dasarnya merupakan penyampaian sebuah pesan, yang berupa informasi, gagasan, ataupun perasaan orang lain [18]. Komunikasi menjadi salah satu aspek kehidupan bagi manusia, di mana dengan adanya komunikasi manusia dapat membentuk hubungan satu dengan yang lain [17]. Komunikasi sangat terkait erat dengan kehidupan seorang manusia di mana dari sejak lahir manusia sudah melakukan komunikasi contohnya sebuah bayi menangis merupakan salah satu bentuk komunikasi pertama yang manusia lakukan [17]. Komunikasi tidak hanya terbatas berbentuk dialog maupun kata-kata yang diucapkan, komunikasi dapat berbentuk interaksi seperti tersenyum atau sikap badan [2].

2.2.2 Telekomunikasi

Telekomunikasi, merupakan sebuah proses pengiriman atau penerimaan infromasi dalam berbagai bentuk seperti isyarat, tulisan, gambar, dan suara melalui sistem radio, optik, kawat, atau sistem elektromagnetik lainnya [3]. Sistem telekomunikasi sendiri mencakup seluruh elemen, mulai dari infrastruktur, perangkat, sarana dan prasarana, serta penyelenggara telekomunikasi. Pertumbuhan industri telekomunikasi didorong karena perkembangan teknologi yang sangat pesat dari pasar telepon seluler. Telekomunikasi pada masa digitalisasi ini telah menjadi salah satu kebutuhan primer untuk manusia. Dengan menggunakan telekomunikasi masyarakat dapat dimudahkan dalam menjalani berbagai aktivitas sehari-hari [5]. Pada masa ini, perusahaan yang bergerak di bidang telekomunikasi seluler tersisa empat antara lain Telkomsel, XL

Axiata, Smartfren, dan Indosat Ooredoo [5]. Pada gambar 2.1 menunjukkan berbagai alasan mengapa masyarakat menggunakan sosial media yang merupakan salah satu bentuk telekomunikasi. Grafik menunjukkan bahwa tujuan utama penggunaan sosial media sebagai sarana komunikasi antar keluarga dan masyarakat.



Gambar 2. 1 Alasan Penggunaan Sosial Media di Indonesia [4]

2.2.3 Data Mining

Data Mining (DM) menurut Pearson merupakan proses analisis data dari berbagai sudut padangan dalam menghasilkan informasi berharga yang dapat membantu perusahaan dalam pengambilan keputusan bisnis yang tepat. Proses dari DM meliputi analisis statistik, penggunaan algoritma *machine learning*, atau teknik lainnya dalam mengidentifikasi pola dan korelasi dalam data. Hasil dari analsisi menggunakan DM dapat digunakan dalam mengindentifikasikan *trend*, risiko, dan peluang. DM merupakan salah satu proses analisis yang sudah banyak di implementasi oleh berbagai bisnis baik untuk mengetahui perilaku pelanggan, memprediksi *trend* industri, maupun membantu mengoptimalkan operasi [19].

2.2.4 Klasifikasi

Klasifikasi merupakan salah satu metode *machine learning* yang digunakan oleh peneliti maupun ahli statistika untuk memprediksi label

kelas dari data yang digunakan. Metode ini merupakan salah satu bentuk analisis data yang digunakan untuk mengelompokkan data ke dalam sebuah kelas berdasarkan atribut tertentu [20]. Klasifikasi merupakan bagian dari *supervised learning* yang berarti metode ini membutuhkan data yang dapat digunakan untuk melatih model terlebih dahulu. Contoh penerapan klasifikasi seperti golongan darah, pendeteksi *spam*, dan *sentiment analysis*. Berikut merupakan jenis klasifikasi yang umumnya ditemukan dalam analisis *machine learning*.

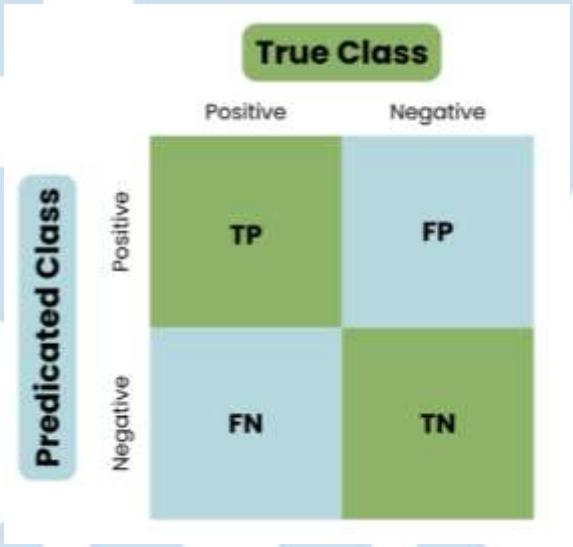
- *Binary Classification* – Merupakan bentuk klasifikasi yang digunakan untuk menangani prediksi antar dua kelas atau label sama halnya seperti bilangan biner yang hanya memiliki 0 atau 1. Contoh *binary classification* seperti sistem pendeteksi spam, dan prediksi *churn* pelanggan. Terdapat beberapa algoritma yang dapat membantu prediksi *binary classification* seperti *K-Nearest Neighbor* dan *Logistic Regression*.

- *Multi–class Classification* – Merupakan bentuk klasifikasi yang digunakan untuk menangani prediksi yang memiliki kelas atau label lebih dari dua. Contoh *multi-class classification* seperti golongan darah, *face recognition*, dan jenis bunga. Algoritma yang dapat membentu masalah *multi-class classification* seperti *Decision Tree* dan *Random Forest*.

2.2.5 Confusion Matrix

Dalam pengembangan model algoritma terdapat beberapa indikator yang dapat menuntukan kualitas model yang digunakan dalam menentukan prediksi yang tepat berdasarkan data yang sudah diberikan. Proses ini dapat dilakukan dengan melihat *confusion matrix* yang dapat memberikan gambaran atau visualisasi terkait performa model prediktif serta melakukan observasi terkait kebenaran kelas yang diprediksi [8]. Pada gambar 2.2 menunjukkan tabel *confusion matrix* yang terdiri dari:

- *True Positive* (TP) – Memiliki arti data *real* yang memiliki hasil positif berhasil diprediksi oleh model dengan hasil positif.
- *False Positive* (FP) – Memiliki arti data *real* yang memiliki hasil negatif yang diprediksi positif oleh model digunakan.
- *True Negative* (TN) – Memiliki arti data *real* yang memiliki hasil negatif berhasil diprediksi oleh model dengan hasil negatif.
- *False Negative* (FN) Memiliki arti data *real* yang memiliki hasil positif yang diprediksi negatif oleh model digunakan.



Gambar 2. 2 Tabel Confusion Matrix [21]

Confusion matrix memiliki metrik yang dapat digunakan untuk evaluasi model prediktif yang dibuat yakni:

- *Accuracy* – Merupakan relasi antara jumlah prediksi yang tepat (TP dan TN) dengan jumlah prediksi yang dibuat oleh model. *Accuracy* dapat mencerminkan seberapa sering model membuat prediksi yang tepat. Tetapi metrik ini kurang sesuai jika kelas atau label yang ingin diprediksi memiliki ukuran data yang tidak seimbang [8].

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

Rumus 2. 1 Rumus *Accuracy*

- *Precision* – Merupakan metrik yang mengukur jumlah prediksi positif yang benar (TP) dan dibandingkan dengan seluruh prediksi positif yang dihasilkan oleh model [8].

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

Rumus 2. 2 Rumus *Precision*

- *Recall* – Merupakan metrik yang mengukur jumlah prediksi positif dengan seluruh contoh positif yang ada dalam data. Metrik ini dapat mewakili bagaimana model dapat menemukan seluruh kejadian positif dengan benar [8].

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

Rumus 2. 3 Rumus *Recall*

- F1 *Score* – Merupakan rata – rata tertimbang antara *precision* dan *recall* dalam model *machine learning*. Metrik ini dapat memberikan gambaran menyeluruh mengenai kinerja model dalam mengidentifikasikan hasil positif yang tepat (TP) serta meminimalkan adanya prediksi positif yang terlewatkan (FN) [8].

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

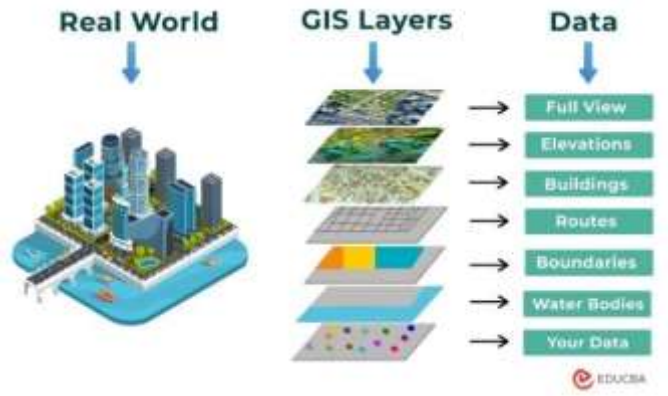
Rumus 2. 4 Rumus *F1-Score*

- *Matthews correlation coefficient* (MCC) – Merupakan evaluasi alternatif yang tidak dipengaruhi oleh kumpulan data yang tidak seimbang, MCC merupakan tingkat *binary classification* yang hanya dapat memberikan nilai tinggi jika model prediktif dapat memprediksi dengan tepat mayoritas contoh data positif dan mayoritas contoh data negatif. Nilai ini akan berada dalam rentang [-1, +1] dengan nilai ekstrem -1 dan +1 dicapai dalam kasus mis-klasifikasi sempurna dan klasifikasi sempurna pada saat yang sama [8].

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Rumus 2. 5 Rumus MCC

2.2.6 *Geographical Information System* (GIS)



Gambar 2. 3 Tampilan Layer GIS [22]

Gambar 2. 4 Peta Persebaran Covid-19[23]

Geographic Information System (GIS) merupakan alat berbasis komputer yang digunakan untuk melakukan pemetaan dan menganalisis setiap kejadian maupun peristiwa yang terjadi di bumi [24]. Teknologi GIS dapat menyimpan, memvisualisasikan, menganalisis, dan menafsirkan data geografis. GIS sendiri dapat berkerja dengan memiliki beberapa *layer* atau lapisan yang dapat membentuk visualisasi berbentuk peta. Teknologi GIS dapat digunakan jika dalam sebuah *dataset* terdiri data *geospasial* atau data geografis seperti *latitude* dan *longitude*. Pada gambar 2.3 menunjukkan contoh *layer* atau lapisan yang membentuk visualisasi

berbasis GIS. Teknologi GIS sudah banyak digunakan oleh berbagai pihak umumnya adalah pembuatan peta persebaran yang menunjukkan wilayah atau lokasi tertentu dan *marker* berwarna yang menunjukkan persebaran yang ingin dilihat. Pada gambar 2.4 menunjukkan contoh peta persebaran Covid-19 di seluruh dunia yang mengimplementasi teknologi GIS.

2.3 Metodologi dan Algoritma *Data Mining*

Proses pembangunan model *machine learning* akan dicapai dengan menggunakan beberapa metodologi yakni, CRISP-DM, KDD, dan SEMMA. Setiap metodologi akan memiliki tahapan serta tujuan yang berbeda dalam siklus pembangunan model. Berikut merupakan perbandingan tahapan – tahapan metodologi *machine learning*:

2.3.1 *Cross-Industry Standard Process for Data Mining*

Cross-Industry Standard Process for Data Mining (CRISP-DM) merupakan salah satu metodologi *data mining* populer yang digunakan dalam menjelaskan siklus *data mining*. Dalam metodologi ini terdapat 6 tahap yang akan menjelaskan proses dalam pembangunan model *machine learning*. Metodologi ini tergolong sangat fleksibel di mana setiap tahap dapat diulang sebanyak yang diperlukan agar model yang dihasilkan dapat lebih optimal [19]. Pada gambar 2.5 menunjukkan *framework* metodologi CRISP-DM.

Gambar 2. 5 *Framework* Metodologi CRISP-DM [25]

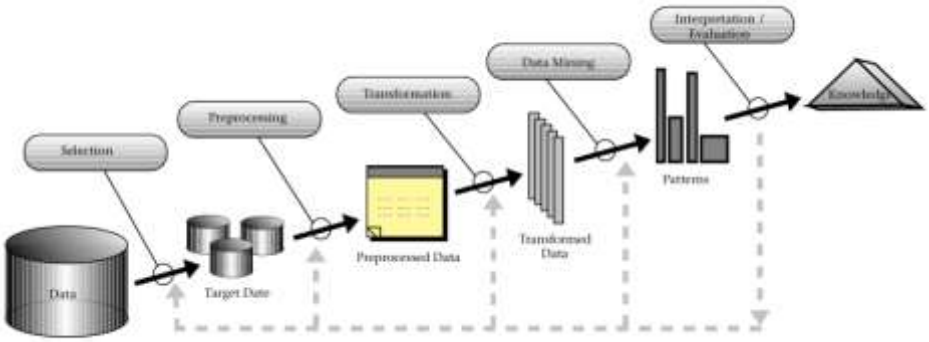
Berikut merupakan tahapan dari *framework* metodologi CRISP-DM [26]:

1. *Business Understanding* – Dalam tahap ini akan melibatkan pemahaman dan persyaratan dari sisi bisnis, yang kemudian akan dibuat menjadi landasan awal dalam pembuatan proyek *data mining*.
2. *Data Understanding* – Dalam tahap ini akan melibatkan pengumpulan serta pemahaman data yang akan digunakan dalam proyek *data mining*. Pemahaman data di sini dapat berupa bentuk penemuan pola, identifikasi kualitas data, dan lain – lain.
3. *Data Preparation* – Dalam tahap ini data yang akan digunakan akan dipersiapkan untuk tahap berikutnya, dalam tahap ini data mentah atau *raw data* akan melakukan proses *transformasi* agar dapat menyesuaikan model yang akan dibuat.
4. *Modeling* – Dalam tahap ini data yang sudah dipersiapkan akan diuji menggunakan teknik pemodelan yang dipilih dan ditetapkan, serta parameter teknis akan disesuaikan untuk membentuk hasil yang optimal.
5. *Evaluation* – Dalam tahap ini, akan dilakukan evaluasi menyeluruh terhadap model yang sudah dibangun, evaluasi ini dilakukan dengan tujuan apakah model yang dibangun sudah sesuai serta memenuhi target sasaran bisnis. Pada akhir tahap ini, keputusan mengenai penggunaan hasil *data mining* perlu dicapai.
6. *Deployment* – Setelah sudah melewati evaluasi tahap terakhir merupakan model yang sudah dibangun dan sesuai dengan sasaran bisnis akan diterapkan. Penerapan ini biasanya akan memiliki bentuk seperti penanaman model ke dalam sistem operasional, melakukan skoring (model akan diterapkan menggunakan data baru), atau penyebaran pengetahuan ke dalam operasional bisnis.

2.3.2 *Knowledge Discovery in Databases* (KDD)

Knowledge Discovery in Databases (KDD) merupakan salah satu metodologi dari *data mining* yang mendapatkan pengetahuan yang bermanfaat dari *database*. Fokus utama dari *framework* ini adalah

menemukan pengetahuan berdasarkan data yang digunakan. Pada gambar 2.6 menunjukkan *framework* metodologi dari KDD [19].

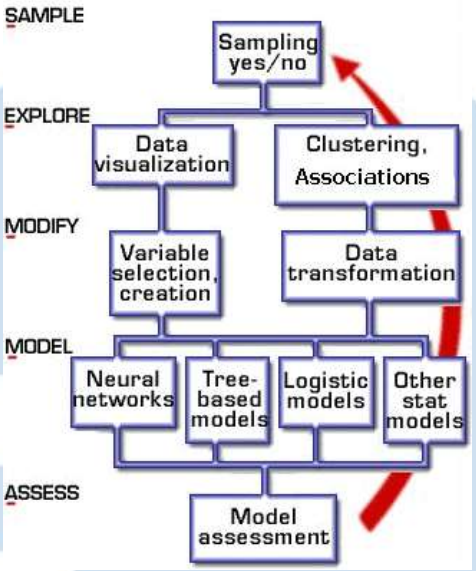


Gambar 2. 6 *Framework* Metodologi KDD [27]

Berikut merupakan tahapan untuk metodologi KDD:

1. *Selection* – Dalam tahap ini meliputi pemilihan serta pemahaman data yang akan digunakan dalam proyek *data mining*.
2. *Preprocessing* – Dalam tahap ini melibatkan pembersihan dan pengelolahan data untuk menghilangkan berbagai *noise* dan inkonsistensi.
3. *Transformation* – Dalam tahap ini merupakan tahap persiapan data untuk *data mining*, data akan melakukan proses transformasi untuk menyesuaikan metode *modeling* yang akan digunakan pada tahap berikutnya.
4. *Data Mining* – Dalam tahap ini data yang sudah dipersiapkan akan diuji menggunakan berbagai teknik pemodelan dengan tujuan untuk mencari pola dari data yang digunakan.
5. *Evaluation* – Dalam tahap ini akan melibatkan evaluasi pola pola yang ditemukan pada tahap sebelumnya serta menyesuaikan dengan tujuan awal dari proyek.
6. *Knowledge* – Merupakan tahap terakhir dalam metodologi KDD di mana seluruh informasi yang didapatkan dari tahap – tahap sebelumnya dapat diterapkan.

2.3.3 *Sample Explore Modify Model Assess* (SEMMA)



Gambar 2. 7 *Framework* Metodologi SEMMA [28]

Sample, Explore, Modify, Assess (SEMMA) merupakan *framework* metodologi *data mining* yang dikembangkan oleh SAS Institude untuk proses *data mining*. *Framework* ini merupakan *toolset* yang digunakan untuk SAS Enterprise Miner dalam menjalankan tugas *data mining*. Pada gambar 2.7 menunjukkan *framework* metodologi dari SEMMA yang terdiri 5 tahapan [29].

Berikut merupakan tahapan – tahapan dari metodologi SEMMA:

1. *Sample* – Dalam tahap ini akan dilakukan pengambilan *sample* data yang akan digunakan untuk pemodelan data. Volume data yang dikumpulkan harus cukup besar agar dapat memuat informasi serta tetap dalam skala yang kecil agar mudah untuk diproses.
2. *Explore* – Dalam tahap ini data yang akan digunakan akan dipelajari atau di *explore* lebih lanjut untuk mencari pola dan relasi di dalam dataset yang digunakan. Tahap ini dilakukan untuk dapat memahami data yang akan digunakan serta sudah bisa mengambil ide dan keputusan yang sesuai. Tahap ini dapat dilakukan dengan berbagai cara tetapi cara yang umum adalah dengan menggunakan visualisasi data.

3. *Modify* – Dalam tahap ini data yang sudah dipahami akan diubah atau mengalami transformasi data agar sesuai dengan model yang akan digunakan pada tahap berikutnya.
4. *Model* – Pada tahap ini data yang sudah disiapkan akan diuji menggunakan berbagai teknik *modeling* yang sudah diterapkan. Tahap ini dilakukan untuk mencapai model yang optimal.
5. *Assess* – Tahap akhir ini melakukan evaluasi keseluruhan untuk seluruh model yang sudah dibangun. Seluruh model yang dibangun akan dibandingankan untuk menemukan model yang paling praktis dan akurat.

2.3.4 Decision Tree

Decision Tree merupakan salah satu algoritma klasifikasi *machine learning*. Algoritma ini mengklasifikasikan datanya menjadi sebuah pohon keputusan, yang terdiri dari berbagai *node* seperti *root node* yang merupakan bagian *node* paling atas [11]. Algoritma ini membagi kumpulan data yang besar menjadi *subset* yang lebih kecil dengan menerapkan seperangkat aturan keputusan [30]. Atribut dari *node* keputusan akan diuji dan setiap hasilnya akan menghasilkan cabang. Setiap cabang ini akan mengarahkan ke *node* lain atau *node* akhir untuk membuat keputusan. *Decision Tree* memiliki struktur seperti *flowchart* di mana setiap *node* akan mewakili nilai atribut, setiap cabang akan mewakili hasil pengujian, dan setiap *leaf* akan mewakili *class* atau *class distribution*.

2.3.5 Random Forest

Random forest merupakan salah satu algoritma yang populer dalam prosedur *machine learning* yang dapat digunakan untuk mengembangkan model prediksi. Algoritma *Random Forest* adalah kumpulan klasifikasi dan pohon regresi yang merupakan model sederhana menggunakan pemisahan biner pada variabel prediktor untuk menentukan hasil dari suatu prediksi. Salah satu algoritma pohon regresi adalah Pohon keputusan yang mudah digunakan dan dapat menawarkan metode intuitif untuk memprediksi hasil yang terbagi menjadi nilai tinggi vs rendah dari sebuah prediktor yang

terkait dengan hasil. Meskipun pohon keputusan menawarkan banyak manfaat, algoritma ini memberikan akurasi yang buruk untuk kumpulan data yang kompleks, terutama kumpulan data yang besar. Dalam klasifikasi *Random Forest* dan pohon regresi dibangun menggunakan dataset pelatihan yang dipilih secara acak dan himpunan bagian acak dari variabel prediktor untuk hasil pemodelan [31]. *Random forest* memiliki beberapa manfaat yang meliputi:

1) Secara umum, memiliki ketelitian yang tinggi
2) Relatif kuat terhadap *noise* dan *outlier*
3) lebih cepat daripada mengantongi dan meningkatkan
4) Sederhana dan mudah untuk diparalelkan

Berikut adalah proses menggunakan algoritma random forest:

1. Ketika sampel *bootstrap* dibangun dengan mengambil data sampel dengan penggantian setiap pohon keputusan, maka sepertiga contoh ditinggalkan.
2. Contoh yang terbengkalai dikenal sebagai *Out of Bags* (OOB) data.
3. Setiap pohon keputusan di hutan memiliki OOB masing-masing dan digunakan untuk memperkirakan kesalahan setiap pohon keputusan
4. Dengan random forest juga dapat menghitung tingkat kepentingan dan perkiraan variabel. Estimasi ini digunakan untuk menghapus dan mengganti nilai dan outlier yang hilang [32].

2.4 *Tools* dan *Software*

2.4.1 Grafana

Grafana merupakan *platform open-source* yang dirancang dalam membantu melakukan visualisasi, pemantauan, serta analisis data. Grafana merupakan salah satu *platform* yang dapat menyediakan layanan visualisasi serta monitoring berbentuk *dashborad* yang interaktif dan menarik [33]. Grafana dalam memberikan layanan pemantauan serta visualisasi sudah mendukung lebih dari 30 *open-source database management system* (DBMS) seperti MySQL, PostgreSQL, Graphite, dll. Grafana menyediakan berbagai macam

layanan visualisasi dan *dashboard*. Selain menyediakan berbagai macam fitur visualisasi dan *monitoring*, Grafana juga memiliki berbagai *plugin* atau ekstensi yang dapat membantu sebuah *programmer* maupun perusahaan dalam melakukan visualisasi, pemantauan, dan analisis data yang lebih mudah dan menarik. Dalam pengembangan sistem *monitoring* berbasis peta, sistem akan dikembangkan menggunakan Grafana sebagai hasil akhir atau *output* yang akan dihasilkan.

2.4.2 Pentaho Data Integration

Pentaho Data Integration (PDI) atau yang dikenal sebagai *kettle*, merupakan *software open – source* yang dapat membantu dalam melakukan proses *extract, transform,* dan *load* (ETL) [34]. *Software* ini memberikan berbagai fitur dan fungsi yang dapat membantu dalam melakukan berbagai macam transformasi data serta memuat data ke dalam *database* atau sebaliknya dalam jumlah volume yang besar. PDI dapat membantu dalam pembuatan program otomatisasi dikarenakan memiliki *schedule run* yang dapat menjalankan program secara otomatis sesuai interval waktu yang ditentukan.

2.4.3 Google Colab

Google Colab merupakan salah satu *software* yang dirancang dan dikembangkan oleh Google di mana untuk *software* ini merupakan *coding environment* untuk bahasa pemrograman Python dengan format *notebook* di mana memiliki kemiripan dengan *Jupyter notebook*. Google Colab menyediakan tiga jenis prosesor untuk keperluan pembelajaran *machine learning* yakni *Central Processing Unit* atau CPU, *Graphic Processing Unit* atau GPU, dan *Tensor Processing Unit* atau TPU. Google Colab menyediakan *environment open source* untuk para programmer ataupun pemula yang ingin mempelajari lebih dalam mengenai Python. Google Colab telah menyediakan berbagai fitur – fitur seperti berbagai *library* Python yang dapat membantu seorang peneliti dalam melakukan preparasi data dan *modeling* data. Selain berbagai fitur untuk membantu sebuah peneliti, programmer, maupun pemula Google Colab juga menyediakan fitur *collaboration* di mana dengan fitur ini seorang pengguna dapat berbagai *coding* secara *online* sehingga akan lebih

mudah dalam melakukan kerja sama antar sesama tim atau mempelajari *coding* orang lain guna untuk mempelajari Python lebih dalam. Google Colab memiliki kelebihan dibandingkan dengan Jupyter Notebook di mana *software* ini tergolong lebih fleksibel dikarenakan dapat langsung terhubung dengan Jupyter Notebook, Google drive, dan Github [35].