BAB II
LANDASAN TEORI

2.1 Penelitian Terkait

Penelitian sebelumnya yang menjadi latar belakang pada penelitian ini, dijabarkan pada tabel dibawah ini :

Tabel 2. 1 Tabel Penelitian Terkait

No.	Nama, Tahun	Judul Penelitian	Metode	Akurasi	Jumlah Dataset
1.	Sri Indra Maiyanti, Des Alwine Zayanti, Yuli Andriani, Bambang Suprihatin, Anita Desiani, Aulia Salsabila, Nyayu Chika Marselina (2023)	Perbandingan Klasifikasi Penyakit Kanker Paru-paru menggunakan *Support Vector Machine* dan K-Nearest Neighbor	*Support Vector Machine (SVM)*	95,16%	309
2.	Muhammad Iqbal Yunan Helmi, Dian Anggraeni, Alfian Futuhul Hadi (2021)	Diagnosis Penderita Penyakit Kanker Paru Menggunakan *Support Vector Machine* dan Naïve Bayes	*Support Vector Machine (SVM)*	90%	80

3.	Dhini Septhya, Kharisma Rahayu, Salsabila Rabbani, Vindi Fitria4, Rahmaddeni, Yuda Irawan, Regiolina Hayami (2023)	Implementasi Algoritma Decision Tree dan *Support Vector Machine* untuk Klasifikasi Penyakit Kanker Paru	*Support Vector Machine (SVM)*	62,3%	1419
4.	Jatnika Fahmi Idris, Rafid Ramadhani, Muhammad Malik Mutoffar (2024)	Klasifikasi Penyakit Kanker Paru Menggunakan Perbandingan Algoritma *Machine Learning*	*Support Vector Machine (SVM)*	55,27%	30000
5.	Emy Haryatmi, Sheila Pramita Hervianti (2021)	Penerapan Algoritma *Support Vector Machine* Untuk Model Prediksi Kelulusan Mahasiswa Tepat Waktu	*Support Vector Machine (SVM)*	94,4%	2181

2.2 Kecerdasan Buatan

Ensiklopedia Britannica mendefinisikan kecerdasan buatan (*Artificial Intelligence*/AI) sebagai cabang ilmu komputer yang dalam merepresentasi pengetahuan lebih banyak menggunakan bentuk simbol-simbol daripada bilangan, dan memproses informasi berdasarkan metode heuristik atau berdasarkan sejumlah aturan. AI tidak sertamerta dapat menggantikan peran manusia dalam industri, tetapi peran AI sebagai pendukung kinerja SDM, oleh karena itu perlunya pengembangan kopentensi oleh SDM yaitu kompentensi yang tidak dapat dilakukan oleh AI. Salah satunya yaitu meningkatkan soft skill SDM.[10]

2.3 Machine Learning

Machine learning (ML) adalah salah satu cabang dari kecerdasan buatan (AI) yang memungkinkan sistem komputer untuk belajar dari data tanpa harus diprogram secara eksplisit. Dengan menggunakan algoritma tertentu, ML mampu menganalisis pola dalam data besar, membuat prediksi, dan memberikan keputusan yang berbasis data. Proses pembelajaran ini melibatkan optimasi parameter model berdasarkan pengalaman atau data historis, sehingga sistem dapat terus meningkatkan kinerjanya seiring waktu. Secara teoritis, *machine learning* dapat dibagi menjadi tiga kategori utama: *supervised learning, unsupervised learning*, dan *reinforcement learning*. Dalam *supervised learning*, model dilatih menggunakan dataset yang sudah diberi label, di mana input dan output diketahui.[11].

2.3.1 Accuracy

Accuracy adalah rasio yang memiliki prediksi nilai benar (nilai positif dan nilai negatif) berdasarkan keseluruhan data. Akurasi dapat menggambarkan keakuratan model klasifikasi yang digunakan. Nilai akurasi dapat diperoleh menggunakan persamaan berikut ini[12]

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

2.3.2 Precision

Precision adalah rasio yang memiliki prediksi nilai benar positif jika dibandingkan dengan keseluruhan hasil yang diprediksi positif. *Precision* dapat menggambarkan keakuratan data yang diinginkan dengan hasil prediksi yang diperoleh model klasifikasi. Nilai Precision dapat diperoleh menggunakan persamaan berikut ini[12].

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

2.3.3 Recall

Recall adalah rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. *Recall* menggambarkan hasil dari model klasifikasi yang digunakan dalam menemukan kembali sebuah informasi. Nilai Recall dapat diperoleh menggunakan persamaan berikut ini [12].

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

2.3.4 F1 Score

F1 Score adalah perhitungan kombinasi dari nilai precision dan nilai recall yang kemudian hasilnya disebut sebagai nilai pengukuran. F1 Score dapat diperoleh menggunakan persamaan berikut ini[12].

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

2.4 Data Mining

Data Mining bertugas untuk menganalisis sejumlah data untuk mencari pola yang khas atau khusus yang belum diketahui sebelumnya seperti kelompok catatan data atau *cluster, anomaly* atau catatan yang tidak biasa. Pola ini kemudian yang dijadikan sebagai rujukan untuk menganalisis data lebih lanjut , misal untuk mengklasifikasi dan memprediksi[13].

Data Mining merupakan proses mengekstrak sejumlah besar data yang sebelumnya tidak diketahui . *Data Mining* juga didefinisikan sebagai bagian dari proses penggalian pengetahuan dari database. Hal ini sering disebut sebagai penemuan pengetahuan dalam keputusan database (KDD) dan bertanggung jawab untuk penyebaran hasil[15].

2.4.1 Pengelompokan *Data Mining*

Data Mining dibagi menjadi beberapa kelompok berdasarkan tugas/pekerjaan yang dapat dilakukan, yaitu :

a. *Descriptive*

Fungsi deskripsi dalam *Data Mining* adalah sebuah fungsi untuk memahami lebih jauh tentang data yang di amati. Dengan melakukan sebuah proses di harap bisa mengetahui perilaku dari sebuah data tersebut. Data tersebut itulah yang nantinya dapat digunakan untuk mengetahui karakteristik dari data yang dimaksud.

b. *Predictive*

Fungsi prediksi merupakan sebuah fungsi bagaimana sebuah proses nantinya akan menemukan pola tertentu dari suatu data. Pola-pola tersebut dapat diketahui dari berbagai variabel-variabel yang ada pada data.

c. *Classification*

Fungsi *Data Mining* ini untuk mengelompokkan atau mengklasifikasikan objek objek dalam data yang memiliki kemiripan. Klasifikasi ini akan membantu proses penemuan definisi dari tiap kesamaan dari suatu kelompok atau kelas.

d. *Association* (Asosiasi)

Asosiasi adalah fungsi *Data Mining* yang tujuannya untuk mengidentifikasi hubungan atau relasi dari setiap data yang ada. Hal ini sejalan dengan istilahnya, asosiasi dipergunakan untuk mengenali dari setiap hubungan atau kejadian yang muncul pada data.

2.5 CRISP-DM *(Cross Industry for Standard Process Data Mining)*

Metode CRISP-DM merupakan metodologi *Data Mining* yang tersusun karena konsorsium perusahaan yang telah didirikan oleh Komisi Eropa di tahun 1996 yang telah diterapkan guna proses standar dalam penerapan *Data Mining*. Tujuan dari metode CRIPS-DM adalah melakukan proses analisis strategi yang digunakan untuk memecahkan masalah penelitian ataupun permasalahan dari sebuah bisnis atau perusahaan. Metode CRISP-DM *(Cross-Industry Standard Process Model for Data Mining)* merupakan penjelasan tentang proses *Data Mining* dengan menggunakan enam tahapan. Metode CRISP-DM memiliki 6 tahapan yaitu *Business Understanding* dengan melakukan *determine business* dan sebuah perencanaan produk. Tahap selanjutnya adalah *Data Understanding* dengan dilakukan pengumpulan data, pendeskripsian data dan eksplor data. Tahap ketiga adalah *Data Preparation* dengan cara *cleansing* data atau pembersihan data kosong atau *null*, *construct* data, *intergrate* data, dan format data. Tahap keempat adalah tahap *Modeling* dimana akan dilakukan pemilihan teknik *modeling, generate test design*, membuat model, dan *assess* model. Tahap selanjutnya yaitu *Evaluation* dengan cara melakukan evaluasi dari data yang telah divisualisasikan atau *modeling* tersebut. Tahap terakhir yaitu *Deployment* dengan memberikan *suggest* untuk strategi kedepannya[16].

2.6 Penyakit Kanker Paru-Paru

Kanker paru-paru merupakan salah satu penyakit dengan tingkat kematian yang tinggi di dunia. Penyakit ini sering kali terdeteksi pada tahap lanjut, sehingga peluang penyembuhannya menjadi lebih kecil. Faktor risiko utama yang berkontribusi terhadap kanker paru-paru meliputi kebiasaan merokok, paparan asap rokok, konsumsi alkohol, kurangnya aktivitas fisik, serta faktor genetik. Oleh karena itu, diperlukan suatu metode yang dapat membantu dalam mendeteksi risiko kanker paru-paru secara lebih dini agar dapat dilakukan tindakan pencegahan dan pengobatan yang tepat.[17]

2.7 Algoritma Yang Digunakan Dalam Penelitian

2.7.1 *Support Vector Machine*

Support Vector Machine (SVM) merupakan metode klasifikasi yang mencari fungsi pemisah *(hyperplane)* terbaik untuk memisahkan dua set data yang berbeda. SVM mencari *margin hyperplane* terbesar antara dua kelas. SVM memiliki performansi yang baik dan mampu menemukan solusi *global optimal*. SVM juga adalah algoritma yang bekerja menggunakan pemetaan *non linier* yang mempunyai tujuan untuk mengubah data pelatihan yang asli ke dalam bentuk dimensi yang jauh lebih tinggi. Kemampuan SVM yang dapat mengolah data berdimensi besar menjadi keunggulan tersendiri dibanding dengan *classsifier* lain[18]. SVM sendiri merupakan algoritma yang kerap digunakan sebagai metode pembelajaran mesin yang di supervisi/diawasi untuk mengklasifikasikan data berdasarkan sentimen positif, negatif, dan netral[18]. Teknik *Support Vector Machine* (SVM) merupakan salah satu teknik pembelajaran learning dengan tingkat akurasi/presisi dan kualitas yang tinggi, menjadikannya algoritma yang sangat populer dibandingkan dengan algoritma lainnya[19]. Ada 2 kernel yang dipakai yaitu, Kernel Linear dan Kernel RBF.

- Kernel Linear

$$K(x_i, x_j) = x_i^\top x_j$$

- Kernel RBF

$$K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right)$$

Kernel linear lebih cocok digunakan pada data yang dapat dipisahkan secara langsung oleh garis lurus, sedangkan kernel RBF lebih unggul dalam menghadapi data yang memiliki distribusi non-linear. Meskipun kernel RBF membutuhkan proses tuning parameter (seperti C dan γ) yang lebih kompleks, namun hasil akurasi dan generalisasi model biasanya lebih baik dibandingkan kernel linear. Oleh karena itu, dalam praktiknya, kernel RBF sering menjadi pilihan utama dalam implementasi SVM untuk permasalahan klasifikasi dengan data yang kompleks dan tidak terstruktur[21].

2.8 Kaggle

Kaggle adalah sebuah platform daring yang sangat dikenal di kalangan ilmuwan data dan mereka yang tertarik dengan *machine learning.* Platform ini menyediakan beragam fasilitas, seperti koleksi data yang melimpah, ajang kompetisi *machine learning*, serta forum komunitas yang ramai. Para pengguna dapat mengikuti kompetisi untuk menguji keahlian mereka, belajar dari sesama ilmuwan data, dan bahkan meraih hadiah. Tak hanya itu, Kaggle juga menyajikan aneka kursus dan tutorial tentang *machine learning*, menjadikannya tempat yang ideal bagi pemula maupun pakar.

2.9 Tools Yang Digunakan Untuk Mengolah Data

2.9.1 Goggle Collab

Google Colab (singkatan dari Google Colaboratory) adalah platform berbasis *cloud computing* yang disediakan oleh Google. Ini memungkinkan pengguna untuk mengeksekusi kode Python dalam lingkungan berbasis *cloud* tanpa perlu menginstal atau mengatur lingkungan lokal mereka sendiri.

2.9.2 Pyhton

Python merupakan salah satu bahasa pemrograman yang banyak digunakan oleh perusahaaan besar maupun para developer untuk mengembangkan berbagai macam aplikasi berbasis desktop, web dan mobile. pemrograman dengan bahasa python sudah banyak dilakukan dengan sifat bahasa pemrogramannya yang *open source* mengakibatkan siapa saja dapat memanfaatkan bahasa pemrograman python. Pada tahun-tahun ini penggunaan bahasa python menjadi semakin relevan dengan analis data.

2.9.3 SpreadSheet

Google Spreadsheet adalah bagian dari Google Workspace dimana terdiri dari GoogleDocs.Dokumen,Spreadsheet,Slide,Formulir. Google Sheets adalah sebuah software atau perangkat lunak berbasis web yang dikembangkan oleh Google, untuk membuat tabel, perhitungan sederhana,atau pengolahan data.