TEKNOLOGI INFORMASI

UNIVERSITAS DARMA PERSADA

BAB II

LANDASAN TEORI

2.1 Kajian terhadap Penelitian Yang Terkait Sebelumnya

Berikut ulasan beberapa penelitian terkait yang menjadi referensi pada penelitian ini: (Diantika et al., 2021) dalam jurnal yang berjudul: "Komparasi Algoritma SVM Dan Naive Bayes Untuk Klasifikasi Kestabilan Jaringan Listrik " pada judul ini penelitian menggunakan metode SVM dan Naïve Bayes untuk membandingkan nilai akhir dari metode tersebut.

Berikut ulasan beberapa penelitian terkait yang menjadi referensi pada penelitian ini: (Sugara & Subekti, 2019) dalam jurnal yang berjudul: "Penerapan Support Vector Machine (SVM) pada Small Dataset untuk Deteksi Dini Gangguan Autisme". Pada penelitian deteksi dini gangguan autisme ini mengusulkan algoritma support vector machine(SVM) untuk memberikan nilai akurasi yang terbaik dengan menggunakan small dataset. Dataset yang dipakai pada pengujian ini sebanyak 67 dengan menghasilkan nilai akurasi yang tertinggi sebesar 85% pada normalized poly kernel. Dua teknik ensemble yaitu Ada Boost dan Bagging juga diusulkan dalam pengujian penelitian deteksi dini gangguan autisme ini untuk meningkatkan kinerja klasifikasi algoritma support vector machine(SVM). Berdasarkan hasil eksperimen yang telah dilakukan menunjukkan bahwa teknik ensemblemenunjukkan performa dapat meningkatkan nilai akurasi. Model SVM dengan poly kernel dan teknik ensemble Bagging menunjukkan nilai akurasi tertinggi yaitu sebesar 91%.

Berikut ulasan beberapa penelitian terkait yang menjadi referensi pada penelitian ini: Hakam Febtadianrano Putro dalam jurnal yang berjudul: "Penerapan

Metode Naive Bayes Untuk Klasifikasi Pelanggan". Penerapan metode naive bayes untuk mengklasifikasikan pelanggan dapat membantu pemilik memberikan bonus terhadap pelanggan berpotensi dan meningkat kualitas yang lebih baik lagi terhadap pelanggan

2.2 Data Mining

Defenisi sederhana data mining adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada didatabase yang besar.

Menurut (Suntoro, 2019) data mining adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan. Sedangkan menurut (Witten, 2016) Data mining adalah proses menganalisa data dari yang berbeda dan menyimpulkannya menjadi informasi atau pengetahuan atau pola yang penting untuk meningkatkan keuntungan, memperkecil biaya pengeluaran, atau bahkan keduanya.

2.3 Dataset

(Suntoro & Kom, n.d.) Data mining tidak pernah lepas dari yang namanya dataset, karena dalam pengolahan data mining, dataset sangat dibutuhkan sebagai objek untuk mendapakan pengetahuan. Dalam terminologi statistik dataset adalah kumpulan dari suatu objek yang mempunyai atribut atau variabel tertentu, di mana untuk setiap objek merupakan individu dari data yang mempunyai sejumlah atribut atau | 6 variabel tersebut. Nama lain dari objek yang sering digunakan adalah record, point, vector, pattern, event, observation, dan case. Sementara itu, baris yang menyatakan objek-objek data dan kolom disebut atribut. Atribut juga dapat disebut dengan variabel, field, fitur atau dimensi.

2.4 CRISP-DM

(Hasanah et al., 2021)Model ini memberikan gambaran tentang siklus hidup proyek data mining. Model ini mempunya 6 Tahapan yaitu:

1. Bussines Understanding

   Ini adalah tahap pertama dalam CRISP-DM dan termasuk bagian yang cukup vital. Pada tahap ini membutuhkan pengetahuan dari objek bisnis, bagaimana membangun atau mendapatkan data, dan bagaimana untuk mencocokan tujuan pemodelan untuk tujuan bisnis sehingga model terbaik dapat dibangun. Kegiatan yang dilakukan antara lain: menentukan tujuan dan persyaratan dengan jelas secara keseluruhan, menerjemahkan tujuan tersebut serta menentukan pembatasan dalam perumusan masalah data mining, dan selanjutnya mempersiapkan strategi awal untuk mencapai tujuan tersebut.

2. Data Understanding

   Secara garis besar untuk memeriksa data, sehingga dapat mengidentifikasi masalah dalam data. Tahap ini memberikan fondasi analitik untuk sebuah penelitian dengan membuat ringkasaan (*summary*) dan mengidentifikasi potensi masalah dalam data. Tahap ini juga harus dilakukan secara cermat dan tidak terburu-buru, seperti pada visualisasi data, yang terkadang *insight*-nya sangat sulit didapat dika dihubungkan dengan *summary data* nya.

3. Data Preparation

   Secara garis besar untuk memperbaiki masalah dalam data, kemudian membuat *variabel derived*. Tahap ini jelas membutuhkan

pemikiran yang cukup matang dan usaha yang cukup tinggi untuk memastikan data tepat untuk algoritma yang digunakan.

4. Modeling

   Secara garis besar untuk membuat model prediktif atau deskriptif. Pada tahap ini dilakukan metode statistika dan *Machine Learning* untuk penentuan terhadap teknik *data mining*, alat bantu *data mining*, dan algoritma *data mining* yang akan diterapkan. Lalu selanjutnya adalah melakukan penerapan teknik dan algoritma data mining tersebut kepada data dengan bantuan alat bantu. Jika diperlukan penyesuaian data terhadap teknik data mining tertentu, dapat kembali ke tahap *data preparation*.

5. Evaluation

   Melakukan interpretasi terhadap hasil dari data mining yang dihasilkan dalam proses pemodelan pada tahap sebelumnya. Evaluasi dilakukan terhadap model yang diterapkan pada tahap sebelumnya dengan tujuan agar model yang ditentukan dapat sesuai dengan tujuan yang ingin dicapai dalam tahap pertama.

6. Deployment

   Tahap *deployment* atau rencana penggunaan model adalah tahap yang paling dihargai dari proses CRISP-DM. Perencanaan untuk *Deployment* dimulai selama *Business Understanding* dan harus menggabungkan tidak hanya bagaimana untuk menghasilkan nilai model, tetapi juga bagaimana mengkonversi skor keputusan, dan bagaimana untuk menggabungkan keputusan dalam sistem operasional.

2.5 Klasifikasi

Klasifikasi merupakan dua bentuk analisa data yang dapat digunakan untuk mengekstrak model yang menggabarkan kelas data atau untuk memprediksi tren data masa depan dan dapat membantu memberikan pemahaman yang lebih baik tentang data secara luas. Proses klasifikasi dibagi menjasi dua tahapan yaitu learning dan test, pada tahap learning data yang diketahui kelas datanya digunakan untuk membangun model dan tahap test dilakukan untuk menguji model yang sudah dibangun untuk mengetahui tingkat akurasi.

2.6 Naïve Bayes

Bayesian classification adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu class. Bayes classificationdidasarkan pada teorema Bayes yang memiliki kemampuan klasifikasi serupa dengandecision tree dan neural network. Bayesian classificationterbukti memiliki akurasi dan kecepatan yang tinggi saat diaplikasikan ke dalam database dengan data yang besar.

NBC merupakan salah satu algoritma klasifikasi yang sederhana namun memiliki kemampuan dan akurasi tinggi.

$$P(Ci|X)=P(X|Ci).P(Ci)P(X)$$

Dengan:

X : data dengan class yang belum diketahui

Ci : hipotesis data X merupakan suatu class spesifik

P(Ci|X) : probabilitas hipotesis Ci berdasarkan kondisi X (posteriori probability)

P(Ci) : probabilitas hipotesis Ci (prior probability)

P(X|Ci) : probabilitas X berdasrkan kondisi pada hipotesis Ci

P(X) : probabilitas dari X

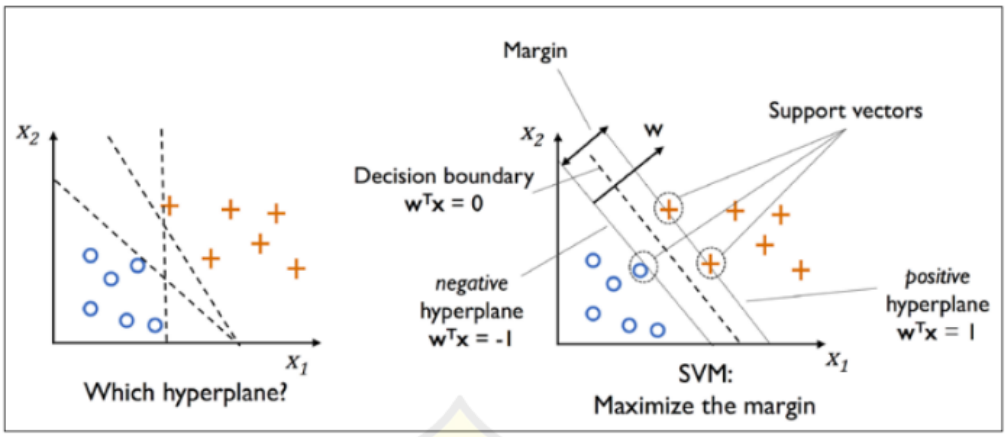
Naïve Bayesian Classifier mengasumsikan bahwa keberadaan sebah atribut (variabel) tidak ada kaitannya dengan beradaan atribut (variabel) yang lain karena asumsi atribut tidak saling terkait (conditionally independent), ditulis dengan rumus: $P(X|Ci)=\sum P(Xk|Cink=1)$

Setelah diperoleh hasil dari seluruh data pada setiap class, maka hasil akhirnya dapat menggunakan rumus.

2.7 SVM (Support Vector Machine)

(Ekonomi et al., n.d.) Support Vector Machine (SVM) merupakan salah satu metode dalam supervised learning yang biasanya digunakan untuk klasifikasi (seperti Support Vector Classification) dan regresi (Support Vector Regression). Dalam pemodelan klasifikasi, SVM memiliki konsep yang lebih matang dan lebih jelas secara matematis dibandingkan dengan teknik-teknik klasifikasi lainnya. SVM juga dapat mengatasi masalah klasifikasi dan regresi dengan linear maupun non linear.

SVM digunakan untuk mencari hyperplane terbaik dengan memaksimalkan jarak antar kelas. Hyperplane adalah sebuah fungsi yang dapat digunakan untuk pemisah antar kelas. Dalam 2-D fungsi yang digunakan untuk klasifikasi antar kelas disebut sebagai line whereas, fungsi yang digunakan untuk klasifikasi antas kelas dalam 3-D disebut plane similarly, sedangan fungsi yang digunakan untuk klasifikasi di dalam ruang kelas dimensi yang lebih tinggi di sebut hyperplane.



Gambar 2.1 Hyperplane yang memisahkan dua kelas positif dan negatif

Hyperplane yang ditemukan SVM diilustrasikan seperti Gambar 2.1 posisinya berada ditengah-tengah antara dua kelas, artinya jarak antara hyperplane dengan objek-objek data berbeda dengan kelas yang berdekatan (terluar) yang diberi tanda bulat kosong dan positif. Dalam SVM objek data terluar yang paling dekat dengan *hyperplane* disebut *support vector*. Objek yang disebut *support vector* paling sulit diklasifikasikan dikarenakan posisi yang hampir tumpang tindih (*overlap*) dengan kelas lain. Mengingat sifatnya yang kritis, hanya *support vector* inilah yang diperhitungkan untuk menemukan *hyperplane* yang paling optimal oleh SVM.