BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Terkait Terdahulu

Tabel 2.1 berikut ini merupakan penelitian terkait dengan penelitian yang sedang dilakukan saat ini :

Tabel 2.1 Penelitian Terkait

No	Nama Penulis	Judul/Tahun Terbit	Uraian
1	Yuniar Kartika , Kokom Komariah , Agus Surip, Riko Saputra , Irfan Ali	Implementasi Algoritma Naïve Bayes Untuk Prediksi Persediaan Barang Rotan KOPERTIP: Jurnal Ilmiah Manajemen Informatika dan Komputer E-ISSN : 2549-9351 Vol. 04 No. 01 Bulan April 2020	Pada era pandemi Covid 19 produksi barang selalu dijaga sesuai dengan kebutuhan dan melihat pada stok rotan. Stok rotan saat ini sedikit terhambat karena terdapat beberapa kebijakan yang dinilai dapat mempengaruhi produksi yaitu Penyekatan antar wilayah. Penelitian ini menitik beratkan pada prediksi penentuan stok rotan agar produksi tetap jalan, bahan baku selalu tersedia. Analisa prediksi stok disesuaikan dari data transaksi penjualan, dari data transakasi penjualan dilakukan analisa menggunakan algoritma naïve bayes. Penitian ini menggunakan data produksi pada tahun 2020 pada CV Jaka Depok Cirebon. Prediksi ini menggunakan aplikasi Rapidminer Versi 9.9 dengan Operator, Retrive, Cross Validation, Naïve

			Bayes, Apply Model dan Performance . Hasil Akurasi pada penlitian ini menunjukan 91.43 % [6].
2	Abdi Rahim Damanik,Sumijan, Gunadi Widi Nurcahyo	Prediksi Tingkat Kepuasan dalam Pembelajaran Daring Menggunakan Algoritma Naïve Bayes Jurnal Sistim Informasi dan Teknologi Vol.3 No. 3 (2021)	Dari hasil pengujian akhir yang dilakukan dari data sampel kuesioner dengan atribut atau indikator komunikasi, suasana pembelajaran, penilaian mahasiswa dan penyampaian materi menggunakan metode Naïve Bayes didapatkan tingkat akurasi sebesar 100% dengan nilai precision sebesar 100% dan nilai recall sebesar 100% [7].
3	Nisa Hanum Harani, Cahyo Prianto	Penerapan algoritma Adaboost guna menentukan pola masuknya calon mahasiswa TRANSFORMTIKA, Vol.18, No.1, July 2020	Penerapan Metode klasifikasi pohon keputusan (decission tree) dan adaboost dapat meningkatkan akurasi hingga 91,35%. Model kombinasi ini dianggap paling akurat jika dibandingkan dengan metode klasifikasi yang hanya menggunakan algoritma pohon keputusan saja (61,4%). Hasil akurasi menunjukan bahwa model yang dihasilkan dapat melakukan prediksi dengan tepat dalam menentukan pola mahasiswa yang akan benar – benar masuk Perguruan Tinggi (PT)[8].
4	Sulaiman Sinaga , Rahmat W.	Penerapan Algoritma Naive Bayes untuk	Berdasarkan pengujian sebanyak 30 data testing yang diolah

	Sembiring , S. Sumarno	Klasifikasi Prediksi Penerimaan Siswa Baru Journal of Machine Learning and Data Analytics (MALDA) Volume 1, No. 1, Januari 2022	menggunakan Rapid Miner, diperoleh tingkat akurasi sebesar 86,6% yaitu 26 siswa diterima dan sebanyak 4 tidak diterima. Sehingga dapat disimpulkan bahwa proses prediksi berupa klasifikasi dengan menggunakan Naive Bayes dapat lebih cepat dan akurat serta menghasilkan tingkat akurasi yang tinggi bila diterapkan untuk mengatasi masalah klasifikasi prediksi jumlah siswa baru pada SMK Anak Bangsa [9].
5	Fazrin Meila Azzahra Sofyan, Apriade Voutama, Yuyun Umaidah	Penerapan Algoritma C4.5 Untuk Prediksi Penyakit Paru-Paru Menggunakan Rapidminer JATI (Jurnal Mahasiswa Teknik Informatika) Vol. 7 No. 2, April 2023	Penelitian ini bertujuan untuk mendapatkan nilai akurasi, recall dan precision dengan menggunakan algoritma C4.5. Adapun data yang digunakan pada penelitian ini diperoleh dari Kaggle yang berisi 30.000 data dengan 11 atribut di dalamnya. RapidMiner digunakan sebagai tools untuk menguji dataset pasien yang digunakan sehingga menghasilkan sebuah pohon keputusan (decision tree) dengan nilai akurasi sebesar 89.77%, recall sebesar 78.61%, dan precision sebesar 100% yang diperoleh dari split data 90% (data training) – 10% (data testing) [2].
6	Devina Larassati , Ati Zaidiah,	Sistem Prediksi Penyakit Jantung Koroner	Berdasarkan hasil dan pembahasan tersebut, maka

	Sarika Afrizal	Menggunakan Metode Naïve Bayes JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika) Volume 07, Nomor 02, Juni 2022	didapatkan kesimpulan bahwa Algoritma Naive Bayes dapat melakukan prediksi penyakit jantung koroner berdasarkan pemeriksaan dini pada pasien. Pada pemilihan pembagian data training dan data testing untuk model yang akan dibangun, hasil akan lebih optimal jika akurasi tertinggi yang dipilih. Berdasarkan evaluasi, hasil tertinggi yang didapatkan sebesar 83,1% [10].
7	Nisa Hanum Harani, Cahyo Prianto	Penerapan algoritma Adaboost guna menentukan pola masuknya calon mahasiswa TRANSFORMTIKA, Vol.18, No.1, July 2020	Penerapan Metode klasifikasi pohon keputusan (decission tree) dan adaboost dapat meningkatkan akurasi hingga 91,35% [8]. Model kombinasi ini dianggap paling akurat jika dibandingkan dengan metode klasifikasi yang hanya menggunakan algoritma pohon keputusan saja (61,4%) . Hasil akurasi menunjukan bahwa model yang dihasilkan dapat melakukan prediksi dengan tepat dalam menentukan pola mahasiswa yang akan benar – benar masuk Perguruan Tinggi (PT)
8	Lidia Pebrianti, Fitrahuda Aulia , Halimatun Nisa , Kana Saputra S	Implementasi Metode Adaboost untuk Mengoptimasi Klasifikasi Penyakit Diabetes dengan Algoritma Naïve Bayes	Adaboost adalah Algoritma Boosting yang paling terkenal, dapat digunakan dengan tujuan untuk meningkatkan keakuratan kinerja pembelajaran Machine Learning Naïve Bayes, sehingga

		JUSTINDO (Jurnal Sistem & Teknologi Informasi Indonesia), Vol. 7, No. 2, Agustus 2022	dapat mengurangi noise dalam kumpulan data yang berukuran besar dengan beberapa kelas atau multi kelas. .Dengan menggunakan split data 60/40 Algoritma Naïve Bayes menghasilkan akurasi sebesar 0.7608. Sedangkan untuk hasil Naïve Bayes yang di boosting dengan menggunakan algoritma Adaboost adalah sebesar 0,7694.

Berdasarkan hasil review beberapa jurnal dengan ini dapat disimpulkan bahwasanya metode yang digunakan untuk beberapa jurnal yang saya review tingkat akurasi cenderung lebih tinggi dengan menggunakan Naïve Bayes memiliki tingkat akurasi yang tinggi, dengan hal ini Adabost dapat digunakan sebagai optimasi level algoritma pada algoritma Naïve Bayes ragar akurasi yang didapatkan lebih tinggi.

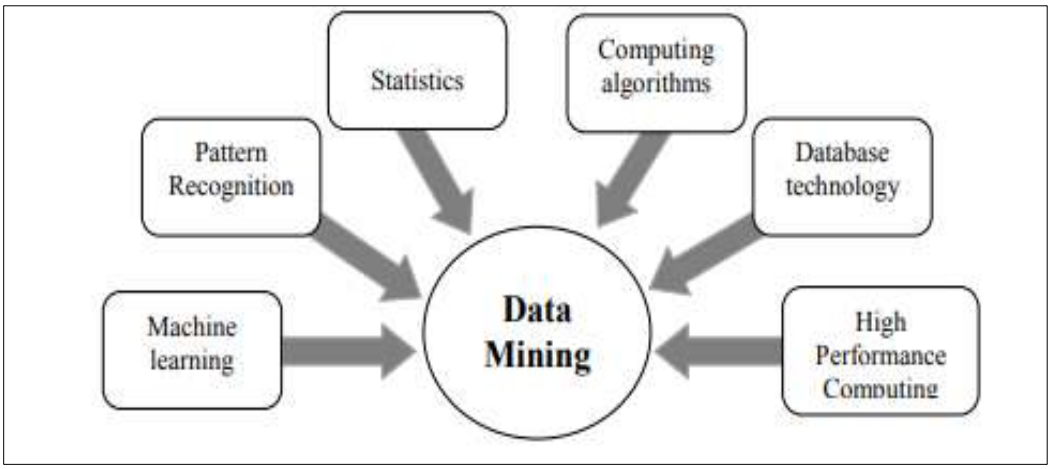
2.2 Penyakit Paru-Paru

Penyakit paru-paru adalah penyakit yang khusus menyerang organ paru-paru. Paru-paru merupakan salah satu organ pernapasan yang berfungsi untuk melakukan repirasi, yaitu mengubah gas (CO2) menjadi gas oksigen (O2) dan air (H2O). sebelum sampai ke paru-paru, udara yang dihirup manusia akan melewati hidung, pangkal laring , kemudian menuju ke kedua bronki utama (bronkus) dan akan disalurkan ke bronki yang paling kecil (bronkioli), selanjutnya udara dimasukan ke dalam jutaan kantong udara (alveoli) yang berada dalam paru-paru. Paruparu memiliki lapisan pelindung (pleura) yang juga berfungsi untuk membantu kontraksi dalam rongga dada. Penyakit paru-paru adalah kondisi paru-paru dimana terjadi peradangan atau pengumpulan cairan (darah atau nana) atau masuknya bakteri, virus atau jamur ke dalam paru-pari yang kemudian menyebabkan paruparu tidak berfungsi dengan baik [11].

2.3 *Data Mining*

Data mining dikenal sejak tahun 1990-an, ketika adanya suatu pekrjaan yang memanfaatkan data menjadi suatu hal yang lebih penting dalam berbagai bidang, seperti marketing dan bisnis, sains, dan teknik, serta seni dan hiburan. Sebagian ahli menyatakan bahwa *data mining* merupakan suatu lankah untuk menganalisis pengetahuan dalam basis data atau biasa disebut *Knowledge Discovery in Database* (KDD) . *Data mining* merupakan proses untuk menemukan pola data dan pengetahuan yang menarik dari kumpulan data yang sangat besar [12].

Data mining, secara sederhana merupakan suatu langkah ekstraksi untuk mendapatkan informasi penting yang sifatnyaimplisit dan belum diketahui. Data miningmempunyai hubungan dengan berbagai bidang seperti statistic, machine learning, *computing algorithms, database technology.* Gambar 2.1 merupakan diagram hubungan *data mining* :



Gambar 2.1. Diagram Hubungan *Data Mining*

Secara sistematis, langkah utama untuk melakukan *data mining* terdiri dari tahap, yaitu sebagai berikut :

1) Ekspolasi Atau Pemrosesan Awal Data

   Eksplorasi atau pemrosesan awal data terdiri dari pembersihan data, normalisasi data, transformasi data, penanganan missing value, reduksi dimensi, pemilihan subset fitur, dan sebagainya.

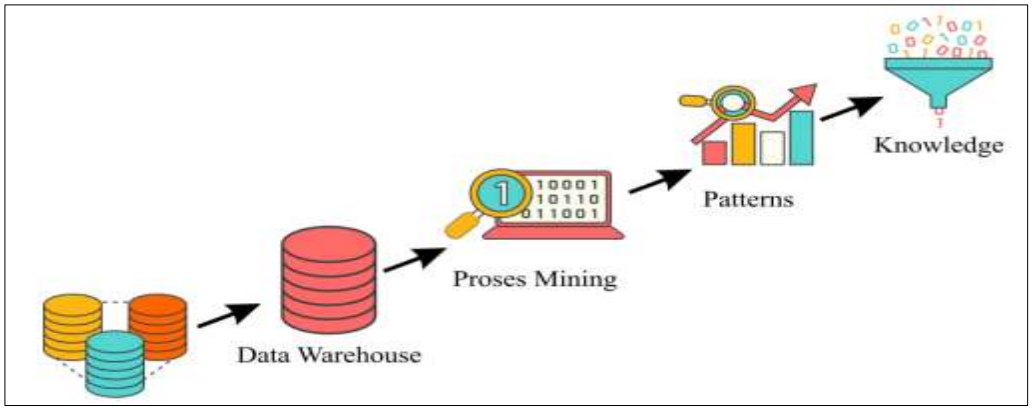
2) Membangun Model Dan Validasi

Membangun model dan validasi, merupakan melakukan analisis dari berbagai model dan memlih model sehingga menghasilkan kinerja yang terbaik. Pembangunan model dilakukan menggunakan metode-metode seperti klasifikasi, regresi,analisis cluster, dan asosiasi.

3) Penerapan

Penerapan dilakukan dengan menerapkan model yang dipilih pada data baru untuk menghasilkan kinerja yang baik pada msalah yang diinvestigasi.

Tahapan proses data mining ada beberapa yang sesuai dengan proses KDD (*Knowledge Discovery in Database*) . Gambar 2.2 merupakan proses KDD (*Knowledge Discovery in Database*) :



Gambar 2.2. Proses KDD (*Knowledge Discovery in Database*)

1. *Cleaning And Integration.*
   a. *Data Cleaning* (Pembersih data)

      Data cleaning (Pembersihan data) adalah proses yang dilakukan untuk menghilangkan noise pada data yang tidak konsisten atau bisa disebut tidak relevan. Data yang diperoleh dari database suatu perusahaan maupun hasil eksperimen yang sudah ada, tidak semuanya memiliki isian yang sempurna misalnya data yang hilang, data yang tidak valid, atau bisa juga hanya sekedar salah ketik. Data yang tidak relevan itu dapat ditangani dengan cara dibuang atau sering disebut dengan proses cleaning. Proses cleaning dapat berpengaruh terhadap performa dari teknik *data mining*.

b. *Data Integration* (Integrasi Data)

Integrasi data merupakan proses penggabungan data dari berbagai database sehingga menjadi satu database baru. Data yang perlukan pada proses *data mining* tidak hanya berasal dari beberapa database.

2. *Selection and Transformation*

a. *Data Selection* (Seleksi Data)

Tidak semua data yang ada didatabase akan dipakai, karena hanya data yang sesuai saja yang akan dianalisis dan diambil dari database.Misalnya pada sebuah kasus market basket analysis yang akan meneliti faktor kecenderungan pelanggan, maka tidak perlu mengambil nama pelanggan, cukup dengan id pelanggan.

b. *Data Trnasformation* (Transformasi Data)

Transformasi data merupakan proses pengubahan data dan penggabungan data ke dalam format tertentu, *data mining* membutuhkan format data khusus sebelum diaplikasikan. Misalnya metode standar seperti analysis asosiasi dan clustering haya bias menerima inputan data yang bersifat katagorial. Karenanya data yang berupa angka numeric apabila mempunyai sifat kontinyu perlu dibagi menjdi beberapa interval. Proses ini sering disebut dengan transformasi data.

3. *Poses Mining*

Proses mining dapat disebut juga sebagai proses penambangan data. Proses mining merupakan proses utama yang menggunakan metode untuk menemukan pengetahuaan beharga yang tersembunyi dari data.

4. *Evaluation and Precentation*

a. Evaluasi Pola *(Pattren Evaluation)*

Evaluasi pola bertugas untuk mengidentifikasi pola-pola yang menarik ke dalam knowledge based yang ditemukan. Pada tahap ini dihasilkan polapola yang khas dari model klasifikasi yang dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai. Bila ternyata hasil yang diperoleh tidak sesuai dengan hipotesa,

terdapat beberapa alternative yang bias diambil seperti menjadikanya umpan baik untuk memprbaiki proses *data mining*, atau mencoba metode *data mining* lain yang lebih sesuai.

b. Presentasi Pengetahuan *(Knowledge Presentation)*

Knowledge presentation merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan atau informasi yang telah digali oleh pengguna. Tahap terakhir dari proses data mining adalah memformulasikan keputusan dari hasil analisis yang didapat.

2.4 *Particle* ADABOOST (Adaptive Boosting)

Adaboost merupakan akronim dari Adaptive Boosting termasuk kedalam Ensemble Methods /Boosting Methods yang sering dipakai. Secara garis besar proses yang dilakukan dalam Adaboost ialah membangun sejumlah weak learners yang tidak memiliki korelasi satu sama lain, lalu kemudian menggabungkan prediksinya. Dalam penerapannya Adaboost dikombinasikan dengan algoritma lain dengan tujuan untuk mengoptimalisasi performa yang dihasilkan. Adaboost $Hk\ (x)$ [13] didefinisikan sebagai:

$$H_k(x) = \sum_{t=1}^{T} \left(\frac{log\ 1}{\beta_t}\right) h_t^k(x) \quad \text{.....(1)}$$

Dimana $ht\ k\ (x)$ merupakan weak learners yang memiliki nilai error terendah , sedangkan βt merupakan bobot dari weak learners tersebut. Premis akhir dalam Adaboost dihasilkan dari kombinasi weak learners yang memiliki nilai suara tertinggi

2.5 Algoritma Naïve Bayes

Algoritma Naïve Bayes menggunakan teknik percabangan matematika dengan mencari peluang terbesar dari kemungkinan dalam klasifikasi berdasarkan frekuensi tiap klasifikasi terhadap data training yang sering di sebut dengan teori probabilistic Adapun rumus perhitungan dari Naïve Bayes adalah sebagai berikut [9]

$$P(X|Y) = \frac{P(Y|X)xP(X)}{P(Y)} \quad \text{……. (2)}$$

Keterangan :

Y = data dengan kelas yang belum diketahui

X = hipotesis data Y merupakan suatu kelas spesifik

P(X| Y) = probabilitas hipotesis X berdasarkan kondisi

Y P(X) = Probabilitas hipotesis

X P(Y| X) = Probabilitas Y berdasarkan kondisi pada hipotesis

X P(Y) = Probabilitas Y

Dalam teori bayes ada beberapa prinsip ketidaktergantungan yang kuat untuk tiap fitur pada sebuah data yang tidak berhubungan dengan adanya fitur lain pada data yang sama. Pada hipotesis Naïve Bayes adalah merupakan label identitas kelas yang telah menjadi target pada pemetaan klasifikasi dengan telah terkait pada korelasi hipotesis, dan bukti tersebut dapat berupa fitur-fitur yang telah menjadi input pada model klasifikasi [15].Selain itu, Algoritma Naïve Bayes Classifier memiliki performa yang sangat baik dalam beberapa kasus klasifikasi teks.

2.6 Confusion Matrix

Matriks konfigurasi adalah tabel yang terdiri dari jumlah baris data uji yang diprediksi benar dan salah dengan model klasifikasi yang digunakan. Tabel Confusion Matrix diperlukan untuk memilih kinerja terbaik dari sebuah model klasifikasi [16]. Confusion matrix adalah matrix 2x2 yang merepresentasikan hasil klasifikasi biner pada suatu dataset. Terdapat beberapa rumus umum yang dapat digunakan untuk menghitung performa klasifikasi. Hasil dari nilai accuracy, precision dan recall bisa ditampilkan dalam persentase [17].

2.6.1 Accuracy (Akurasi)

Akurasi adalah salah satu metrik untuk mengevaluasi model klasifikasi. Secara informal, akurasi adalah sebagian kecil dari prediksi model kami yang benar. Secara formal, akurasi memiliki definisi sebagai berikut: Untuk klasifikasi biner, akurasi juga dapat dihitung dalam hal positif dan negatif sebagai berikut:

$$\text{Akurasi} = \frac{\text{Number of Correect Prediction}}{\text{Total Number of Prediction}} \quad \ldots\ldots\ldots\ldots\ldots (3)$$

Untuk klasifikasi biner, akurasi juga dapat dihitung dalam hal positif dan negatif sebagai berikut:

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN}$$

Dimana TP = True Positif

TN = True Negatif

FP = False Positif

FN = False Negatif

2.6.2 Precision

Precision dalam Confusion Matrix didefinisikan sebagai rasio item terkait yang dipilih dengan semua item yang dipilih. Akurasi adalah kemungkinan bahwa item yang dipilih terkait. Dapat diartikan sebagai kecocokan antara permintaan informasi dan respons terhadap permintaan itu [18]

2.6.3 Recal

Recall adalah proporsi jumlah dokumen teks yang relevan terkendali diantara semua dokumen teks relevan yang ada pada koleksi [17].Recall merupakan probabilitas bahwa suatu item yang relevan akan dipilih. Recall dapat dihitung dengan jumlah rekomendasi yang relevan yang dipilih oleh user dibagi dengan jumlah semua r nendasi yang relevan baik dipilih maupun rekomendasi yang tidak terpilih [18]

2.7 Kurva ROC dan AUC

Dalam Machine Learning, pengukuran kinerja adalah tugas penting. Jadi dalam masalah klasifikasi, kita dapat mengandalkan Kurva AUC - ROC. Ketika kita perlu memeriksa atau memvisualisasikan kinerja masalah klasifikasi multikelas, kita menggunakan kurva AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) [19]