BAB II

LANDASAN TEORI

2.1. Penelitian Terdahulu

Pada Tabel 2.1 dibawah ini merupakan penelitian terdahulu yang terkait dan menjadi acuan pada penelitian ini.

No	Author & Tahun	Tujuan	Metode	Sampel	Hasil
1	Predicting students' academic performance using a modified kNN algorithm. Moohanad Jawthari and Veronika Stoffov, 2021	Algoritma K-Nearest Neighbour dimodifikasi untuk menghitung nilai jarak kategorikal (nominal) variabel data kinerja akademik siswa tanpa mengkodekannya.	K-Nearest neighbors	Dataset 480 catatan siswa dengan 16 fitur.	Menunjukkan Algoritma yang diusulkan memiliki akurasi 14% lebih baik daripada yang standar, dan tidak sensitif terhadap outlier.
2	A Review of Students' Graduation Classification: A Comparison of *Naive Bayes* Classifier and K Nearest Neighbour Tuhamah Fauziastuti, Lilis	Tujuan utama dari makalah ini adalah untuk membandingkan pencapaian kedua Algoritma (NBC dan KNN) terhadap klasifikasi kelulusan siswa.	Penelitian ini didasarkan pada teknik data mining dengan mengimplementasikan dua Algoritma umum yaitu *Naive Bayes* Classifier dan K-Nearest	kelulusan siswa Dan CGPA	Hasil penelitian menunjukkan nilai akurasi NBC dan KNN, NBC memiliki skor yang lebih tinggi yaitu 91,10% daripada KNN 85,15%.

	Aslihah Rakhman,2021	Makalah ini juga berfokus untuk mengidentifikasi variabel yang paling penting untuk memprediksi kinerja siswa.	Neighbor karena mengklasifikasikan kelulusan siswa (tepat waktu dan lembur).		
3	Application of Data Mining Classification Method for Student Graduation Prediction Using K-Nearest Neighbor (K-NN) Algorithm. Mohammad Imron ,Satia Angga Kusumah , 2018	Penelitian ini bertujuan untuk mengetahui tingkat akurasi yang telah disampaikan oleh Algoritma K-Nearest Neighbor (K-NN) dalam memprediksi tingkat kelulusan mahasiswa di Stmik Amikom Purwokerto.	Metode K-Nearest Neighbor.	Data yang digunakan berasal dari data siswa, data nilai siswa, dan data kelulusan siswa tahun ajaran 2010-2012 sebanyak 2.189 record. Atribut yang digunakan adalah jenis kelamin, tingkat studi, program studi untuk Indeks Prestasi 1 sampai 6	Hasil penelitian memilikki tingkat akurasi yang tinggi yakni sebesar 89,04%
4	Machine Learning Algorithms for Student Employability Prediction Using R G. Vadivu*, K.Sornalakshmi,2017.	Memprediksi keterampilan kerja berdasarkan kinerja reguler mereka.	Algoritma Pembelajaran Mesin K nearest Neighbour dan Naïve Bayes	Algoritma diterapkan pada kumpulan data 250 siswa dengan 59 atribut.	Akurasi yang diperoleh setelah analisis untuk KNN adalah 95,33% dan untuk naive Bayes adalah 97,67%.

5	Identification of Student Academic Performance using the KNN Algorithm Aldi Nugroho, Osvaldo Richie Riady, Alexander Calvin, Derwin Suhartono, 2020	untuk memantau dan mengevaluasi kegiatan belajar mengajar menggunakan klasifikasi Algoritma KNN dan yang lebih spesifik adalah mengukur perkembangan akademik mahasiswa di universitas, sehingga dapat menyeimbangkan kualitas proses penyampaian materi yang diterima secara merata oleh setiap mahasiswa sesuai dengan kemampuannya.	Metode yang kami gunakan adalah Algoritma KNN yang sama	Dataset yang digunakan ialah data siswa yang tidak lulus.	Pada percobaan menggunakan metode KNN hasilnya terlihat jelas dan menunjukkan akurasi yang cukup baik. Dari percobaan disimpulkan bahwa jika ada sedikit keluhan dari satu mahasiswa dapat meminimalkan tingkat mahasiswa non-lulusan di universitas dan pada akhirnya menghasilkan akreditasi yang baik.
6	Predict Academic Performance of Students using an K Nearest Neighbour Algorithm Case Study, MATLAB Course	Penelitian mencoba untuk mencapai tujuan berikut: Untuk mengetahui apakah pilihan terbaik untuk algoritma	Teknik klasifikasi direpresentasikan dalam Algoritma K-nearest neighbor (KNN)	Pengumpulan data mahasiswi Jurusan Kimia Fakultas Ilmu Pendidikan Afif kehadiran mata kuliah MATLAB ,	penelitian ini mencapai kesimpulan sebagai berikut: 1) Pilihan terbaik untuk nilai (k) Dalam Algoritma mengarah

	Noha Hassan Osman Ragab, Dr Saif Eldin Fattoh,2015	mengklaim prediktabilitas tingkat siswa yang berbeda. Untuk mengetahui peran prediktabilitas kinerja akademik siswa perempuan dalam kurikulum untuk membantu meningkatkan metode pengajaran dan output pendidikan.		UTS I , UTS II dan Indeks Prestasi Kumulatif (IPK) tahun 1435 -1436 hasil karya model dengan memprediksi akademik kinerja mahasiswi tahun 1437 semester pertama dan memasukkan data pada program (Microsoft Excel 2010) kemudian data disimpan pada jenis file (CSV) untuk dimasukkan ke dalam program weka dan pembentukan model prediktabilitas untuk digunakan dalam peramalan prestasi akademik mahasiswi pada mata kuliah MATLAB semester 1 tahun 1437.	pada memprediksi lebih dari jumlah siswa perempuan yang mungkin untuk mengetahui kinerja akademik mereka. Temukan pada pilihan nilai kami,(k=1) Jumlah total siswa perempuan adalah 28 siswa perempuan yang diperkirakan hanya 26 siswa. Bila nilai (K=2) memprediksi angka 21 hanya meminta bila nilai (K=3) diprediksi untuk semua siswa perempuan. Pengetahuan tentang prestasi akademik mahasiswi dalam suatu kurikulum

					membantu staf pengajar untuk mengetahui titik kelemahan dan kelebihan mahasiswi, juga membantu meningkatkan metode pengajaran dan output pendidikan. Prediktabilitas kursus akan membantu untuk menilai kinerja anggota staf pengajar dan pengetahuan tentang kegagalan dan negatif dan pekerjaan untuk mengatasi sebelum akhir semester.
7	Predicting Students'Academic Performance Using Naïve Bayes. Abdullah Baz, Fatima Alshareef, Ebtihal	Tujuan dari penelitian kami adalah untuk memprediksi kinerja akademik mahasiswa di Universitas Umm Al-Qura dengan	Algoritma klasifikasi yang disebut *Naïve Bayes* digunakan pada dataset dengan menggunakan alat WEKA.	Data dikumpulkan dari database Universitas Umm Al-Qura. Dataset ini terdiri dari 138 catatan mahasiswa yang lulus	Hasil yang dicapai menunjukkan bahwa *Naïve Bayes* dapat digunakan untuk memprediksi prestasi akademik siswa pada

	Alshareef, Hosam Alhakami, Tahani Alsubait, 2017	menggunakan metode *Naive Bayes*, salah satu Algoritma klasifikasi data mining yang paling terkenal. Pengklasifikasi ini membantu memprediksi IPK akhir siswa pada tahap awal berdasarkan nilai mata kuliah di tahun pertama.		dari Sekolah Tinggi Komputer dan Sistem Informasi pada tahun 2019, terkait dengan 13 atribut termasuk ID mahasiswa, jenis kelamin, nilai delapan mata kuliah, IPK semester pertama dan kedua di tahun pertama dan IPK akhir.	tahap awal di tahun pertama dengan akurasi 72,46%.
8	Students performance prediction using KNN and Naïve Bayesian. Ihsan A. Abu Amra, Ashraf Yunis Maghari, 2017	Makalah ini mengusulkan model prediksi kinerja siswa menggunakan KNN dan *Naïve Bayes* sebagai teknik klasifikasi yang diterapkan pada kumpulan data untuk sertifikat Umum sekunder, yang dikumpulkan dari kementerian pendidikan di	menerapkan dua Algoritma klasifikasi: KNN dan Naïve Bayes	Menurut statistik kementerian pendidikan di Jalur Gaza untuk tahun 2015, siswa yang hadir untuk mendapatkan Sertifikat Umum Menengah (SGC) adalah 33.294 tetapi 7809 di antaranya tidak berhasil. Jumlah ini mewakili 27% dari total jumlah. Apalagi mewakili	Dalam penelitian kami yang disajikan, dengan membandingkan tiga parameter evaluasi (akurasi, Recall dan Precision) untuk dua Algoritma KNN dan Naïve Bayes, Algoritma NaïveBayes memiliki akurasi tertinggi 93,17% yang berarti hubungan yang kuat

		Jalur Gaza.		sekitar 55% dari total jumlah siswa di beberapa sekolah.	antara fitur yang mempengaruhi kinerja siswa, dan akan membantu untuk prediksi siswa kinerja untuk tahun depan. *Naïve Bayes* lebih baik daripada KNN, yang berarti hubungan yang kuat antara fitur yang mempengaruhi kinerja siswa, dan itu akan membantu untuk prediksi kinerja siswa. Terkadang, KNN akan lebih baik daripada *Naïve Bayes* untuk kumpulan data lain dan IDE yang berbeda. Sebagai pekerjaan di masa depan, lebih banyak Algoritma klasifikasi dapat diterapkan pada kumpulan

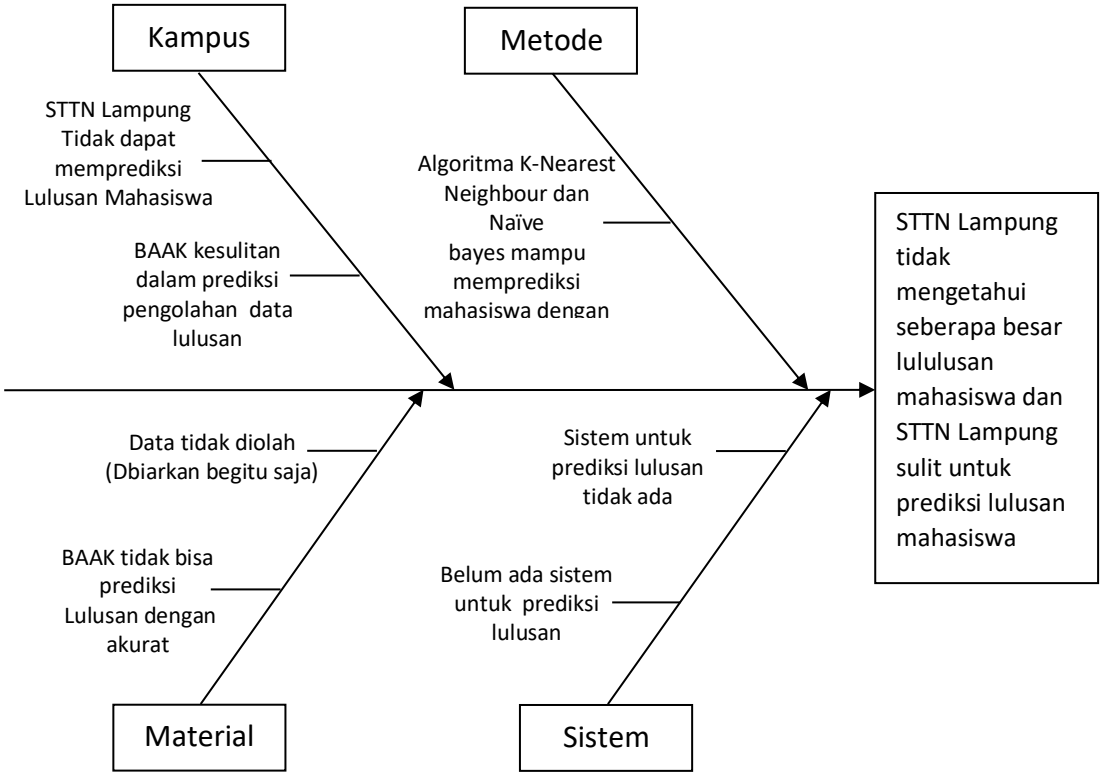
					data pendidikan yang berbeda.
9	Text Classification for Student Data Set using*Naive Bayes* Classifier and KNN Classifier. Rajeswari R.P, Kavitha Juliet, Dr.Aradhana, 2017	Untuk menekankan kinerja dan akurasi pengklasifikasi ini menggunakan penambang cepat untuk Kumpulan Data Siswa.	Pengklasifikasi *Naïve Bayes* dan pengklasifikasi K-Nearest Neighbor	lembar excel siswa dengan atribut sebagai berikut. USN, Nama Jurusan, Nama Perguruan Tinggi, Jenis Kelamin, Usia, Tahun Studi, AGPY, Beasiswa	Percobaan yang dilakukan menunjukkan bahwa pengklasifikasi Naives Bayes merupakan pengklasifikasi yang baik dengan akurasi 66,67 dibandingkan pengklasifikasi KNN dengan 38,89.
10	Variable Selection to Determine Majors of Student using K-Nearest Neighbor and *Naïve Bayes* Classifier Algorithm Mustakim Mustakim, Zarkasih Zarkasih, Petir Papilo, Zaitun Zaitun, 2019	Banyaknya variabel yang digunakan dalam seleksi,menyebabkan beberapa kelemahan di antaranya, seperti kompleksitas variabel, infisiensi variable dan adanya beberapa variabel yang hanya sebagai tambahan tanpa memiliki kontribusi yang	Akurasi Matriks yang diukur menggunakan Metode K-Nearest Neighbor dan Naïve Bayes Classifier (NBC) .	Ada banyak variabel yang harus dipertimbangkan untuk menentukan jurusan siswa, seperti: Jenis Kelamin, Minat, Intelligence Quotient (IQ); Empat mata pelajaran di SMP (JHS), rata-rata nilai SMP, nilai pendaftaran empat mata pelajaran, dan rata-rata	Hasil eksperimen menunjukkan bahwa kombinasi variable yang menghasilkan akurasi terbaik adalah percobaan ke-9 dan ke-10 dengan matrikulasi variabel,minat dan IQ, dan akurasi 96,77% dari K-NN juga 98,38% dari NBC. Dengan menggabungkan kedua Algoritma, 99,87%

		signifikan. Penelitian ini bertujuan untuk mengurangi jumlah variabel tersebut agar dapat menjadi lebih mudah untuk dianalisis dan diterapkan.		tingkat matrikulasi.	akurasi maksimum diperoleh dari ketiga variabel tersebut. Baru Informasi yang dapat diperoleh dari penelitian ini adalah hanya ada tiga hal penting variabel untuk menentukan penempatan jurusan di SMA, Nilai Rata-rata Matrikulasi, Minat dan IQ diikuti oleh empat variabel pendukung seperti nilai Matematika, Fisika,Bahasa Inggris dan Ekonomi dalam Matrikulasi.

Berdasarkan tabel penelitian di atas, maka yang menjadi keunikan dari penelitian ini terletak pada model penerapan metodologi, kasus dan hasil penelitian yang berbeda. Dimana pada tabel penelitian diatas menggunakan Algoritma *Naive Bayes* dan *K-Nearest Neighbor*s untuk memprediksi kelulusan, sedangkan pada penelitian ini Algoritma *Naive Bayes* dan *K-Nearest Neighbor*s digunakan untuk memprediksi kelulusan mahasiswa pada Sekolah Tinggi Teknologi Nusantara Lampung.

2.2. Diagram *Fishbone*

Diagram *Ishikawa* (tulang ikan/diagram sebab akibat) adalah diagram yang digunakan untuk mengidentifikasi masalah (Kaoru Ishikawa, 1968). Pada diagram ini merupakan diagram yang menunjukkan penyebab dari suatu peristiwa tertentu. pada umumnya diagram digunakan untuk mengidentifikasi faktor-faktor yang mempengaruhi suatu peristiwa. Gambar 2.2 diagram *Ishikawa* pada penelitian.



Gambar 2.1. *Ishikawa* Diagram Sistem

Dari gambar di atas dapat dilihat bagaimana sistem memprediksi kelulusan siswa Sekolah Tinggi Teknologi Nusantara Lampung, pada diagram *fishbone* dibagi menjadi dua bagian, yaitu kepala dan tulang. Ada empat (Bone) aspek, yaitu materi, metode, kampus dan sistem. Aspek inilah yang dibutuhkan untuk

membuat sistem bekerja yaitu memprediksi kelulusan mahasiswa algoritma *K-Nearest Neighbors* dan algoritma *Naive Bayes*. Metode tersebut sesuai dengan kebutuhan spesifik dari proses dalam bentuk *ekstraksi fitur*, data *preprocessing*. Kampus sebagai pengguna yang melakukan proses interaksi dengan system, yaitu menentukan dataset yang dapat digunakan dalam sistem sehingga data dapat dilatih dan diuji dengan baik. Metode tersebut merupakan kebutuhan proses khusus berupa ekstraksi ciri validasi hasil pelatihan dengan 10 *fold cross validation*, dan *preprocessing* data.

2.3. *Machine Learning*

Bagian dari bidang ilmu yang berhubungan dengan statistik, kecerdasan buatan dan ilmu komputer yang dapat memprediksi hasil analisis dan studi data adalah pembelajaran mesin (Andreas C.Muller & Sarah Guido, 2017). Proses *Machine Learning* sama dengan data mining. Namun dalam pembelajaran mesin, Mereka mencari model, dengan menggunakan data untuk meningkatkan pemahaman tentang program itu sendiri. (Widodo Budiharto, 2016). Manfaat penggunaan *machine learning* adalah bagaimana kita merencankan yang kita lakukan di masa depan dengan mengolah data yang kita ubah (Judith Hurwitz & Daniel Kirsch, 2018).

2.3.1. Jenis-Jenis Algoritma Pembelajaran *Machine Learning*

Jenis-jenis algoritma pembelajaran mesin adalah sebagai berikut:

1). *Supervised Learning*

Penerapan konsep mengajar dari guru ke siswa merupakan pengertian *supervised learning*. Penerapan algoritma dalam menyediakan data pelatihan yang akurat, dapat dilakukan pengukuran dengan setiap iterasi dan perbedaan dimana nilai prediksi dan nilai yang diinginkan dapat diketahui (Giuseppe Bonaccorso, 2017). Pembelajaran yang lebih mudah dimana mesin belajar dari contoh baik dan buruk merupakan implementasi *supervised learning*. Beberapa kasus dan kemudian menyimpan sendiri ketika *output* berjalan terus menerus dapat dipelajari oleh mesin (Ethem Alpaydin, 2010). Algoritma dilatih berdasarkan contoh yang telah diproses sebelumnya dan kinerja algoritma dinilai berdasarkan hasil data uji. Seringkali pola di jadikan menjadi bagian data urut dan tidak bisa dicari pada data latih besar. Jika model hanya mewakili model yang sudah ada pada *subset* pelatihan, yang berarti model yang digunakan untuk data pelatihan tidak dapat diterapkan pada data pelatihan yang lebih besar dan tidak diketahui maka terjadi *overfitting* (Judith Hurwitz & Daniel Kirsch, 2018).

2) *Unsupervised Learning*

Pemetaan dilakukan dari *input* ke *output* di mana nilai yang benar diberikan oleh pengguna merupakan pengertian *Unsupervised Learning*., pengguna hanya memasukkan data pelatihan tanpa melabeli data pelatihan pada *unsupervised learning* (Ethem Alpaydin, 2010). Kesalahan pengukuran dikarenakan *Unsupervised Learning* mampu belajar sendiri

tanpa pengawasan. Metode ini sangat efektif untuk mengelompokkan objek yang sama. Pembelajaran pada model ini didasarkan pada bagaimana *item* dikelompokkan berdasarkan kesamaan (Giuseppe Bonaccorso, 2017). *Unsupervised Learning* mencocokkan data dalam kelompok karakteristik. Data yang tidak memilikki label membuat nilai parameter pada klasifikasi data. Hasil ketika data yang cukup dapat diberikan oleh *Unsupervised Learning*. Dengan demikian data mana yang dianalisis, pengguna tidak mengetahui. Dalam mengelola data pelatihan yang besar dan tidak ditandai *Unsupersived Learning* sangat berguna (Judith Hurwitz & Daniel Kirsch, 2018).

3) *Reinforcement Learning*

Memimpin sistem pengambilan keputusan di mana mesin melakukan tindakan dan menerima penghargaan atau hukuman atas tindakan yang diperlukan untuk memecahkan masalah merupakan metode pembelajaran *reinforcement learning*. Pembelajaran penguatan menggunakan sistem coba-coba untuk mencapai nilai dan jangkauan maksimal pada titik maksimalnya (Ethem Alpaydin,2014). *Reinforcement Learning* adalah pemodelan pembelajaran yang berdasarkan perilaku. Umpan balik dari analisis data sehingga pengguna cukup membimbing mereka untuk mencapai hasil terbaik yang diterima algoritma dapat membedakan jenis pembelajaran lainnya dari pembelajaran penguatan, bahwa sampel data tidak memerlukan pembelajaran penguatan. Sistem dijalankan dengan *trial* and *error*. Pada bidang aplikasi robotika dan game, sistem mobil pintar

adalah contoh implementasi pembelajaran penguatan. (Judith Hurwitz & Daniel Kirsch, 2018).

4) *Neural Network* dan *Deep Learning*

Pada *Deep Learning* algoritma pembelajaran mesin yang menggabungkan jaringan saraf dalam pembelajaran data interaktif. Bila diterapkan pada model pelatihan dari data yang tidak beraturan *deep learning* sangat berguna. Sehingga pelatihan menghasilkan abstraksi dan masalah yang tidak jelas dikarenakan pada *deep learning* itu bekerja seperti otak manusia. Teknik pembelajaran mesin ini menggunakan *hierarki* jaringan saraf yang diperoleh dari kombinasi *supervised learning* dan *unsupervised learning*. *Deep Learning* umumnya disebut *sub* bagian dari pembelajaran mesin. Tiga lapisan dimilikki jaringan saraf, yaitu lapisan *input*, lapisan *output* dan lapisan tersembunyi. Pada *input layer* data disisipkan kemudian dikelompokkan sesuai dengan bobot yang diterapkan pada setiap *node* dan diproses pada *hidden layer* juga *output layer*. (Judith Hurwitz & Daniel Kirsch, 2018).

2.3.2. Jenis-Jenis Algoritma *Machine Learning*

1) *Bayesian*

Label pada data model akan terlihat dan i*ndependensi* data yang diberikan *Bayesian*. *Bayesian* memiliki banyak fokus dalam pemodelan data hal ini sangat efektif bila diterapkan pada sejumlah kecil data pelatihan.

2) *Clustering*

Teknik untuk memahami suatu objek dengan parameter yang sama sehingga dapat dikelompokkan merupakan pengertian *Clustering*. Semua objek dapat dikelompokkan berdasarkan kesamaannya. Algoritma yang menerapkan pembelajaran *unsupervised learning* biasanya menggunakan *Clustering*.

3) *Classification*

Klasifikasi adalah pengelompokan data dimana data yang digunakan memiliki target kelas atau label. Dengan demikian, klasifikasi algoritma pemecahan masalah diklasifikasikan sebagai *supervised learning* atau pembelajaran yang diawasi. Tujuan pembelajaran terawasi adalah agar label data atau tujuan bertindak sebagai "pengawas". atau guru yang mengawasi proses pembelajaran untuk mencapai tingkat ketelitian atau ketepatan tertentu.

4) *Decision* Tree

Struktur cabang yang menggambarkan hasil keputusan menggunakan *decisionn tree* dapat digunakan untuk proses kemungkinan hasil keputusan. Setiap simpul dari pohon keputusan diwakili oleh hasil yang mungkin. Dalam menetapkan *node* berdasarkan kesamaan hasil yang diperoleh menggunakan persentase.

5) *Reduksi* Dimensi

Data yang tidak digunakan pada tahap analisis dibantu *Reduksi* Dimensi sistem. Menghilangkan *redundansi* data dan data lain yang tidak perlu merupakan kelompok algoritma ini. Saat menganalisis data dalam aplikasi *Internet of Things* (IoT) berguna dalam*Minimizing*.

6) Berbasis Contoh *(lazy learners)*

Ketika memeringkat data berdasarkan kesamaan data pelatihan, algoritma berbasis sampel biasanya digunakan. Karena tidak ada proses pelatihan bisa disebut juga sebagai *lazy learner* pada jenis algoritma ini. Algoritma ini biasanya mengklasifikasikan kedalam kategori data baru berdasarkan kesamaan dalam data pelatihan serta mencocokkan data dengan data pelatihan.

7) Jaringan Saraf Tiruan dan *Deep Learning*

Terlihat seperti saraf manusia dalam menangani suatu masalah dan menggunakan lapisan yang saling terkait dalam analisis data dalah jaringan saraf tiruan. Jika ada lebih dari satu lapisan tersembunyi, itu disebut pembelajaran mendalam.

8) *Regresi Linear*

Algoritma *regresi* umumnya digunakan dalam analisis statistik. Algoritma *regresi* dapat digunakan untuk memprediksi nilai data berdasarkan nilai *historis* juga menganalisis hubungan antara model serta data sehingga

dapat mengukur korelasi antar variabel dalam data latih. (Judith Hurwitz dan Daniel Kirsch,2018).

2.4. Data *Mining*

Ekstraksi atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar merupakan pengertian data *mining* (Davies, 2004). Data mining digunakan untuk menemukan pola yang menarik dari data dalam jumlah besar, pada database data tersebut dapat disimpan, data *warehouse* atau penyimpanan informasi lainnya (Han, 2006). Penambangan data terkait dengan bidang ilmu lain, seperti sistem basis data, pergudangan data, statistik, pembelajaran mesin, temu kembali informasi, dan komputasi kinerja tinggi. *Neural network, pattern recognition, spatial* data *analysis, image database, signal processing* mendukung ilmu-ilmu pada data *mining*. Data *mining* dapat dibagi menjadi beberapa tahap yang diilustrasikan pada gambar sebagai rangkaian proses tahapan data *mining*. *Fase-fase* ini bersifat *interaktif*, pengguna terlibat secara langsung atau melalui *basis* pengetahuan. Ada 7 tahapan data mining, yaitu:

1) Pembersihan data (Data *Cleaning*) adalah proses menghilangkan data yang tidak konsisten atau tidak *relevan* atau *noise*. Data yang diperoleh dari database perusahaan maupun dari hasil *eksperimen*, memiliki entri yang tidak sempurna seperti adanya data yang hilang, data yang tidak *valid*. *Hipotesis* penambangan data yang membuat atribut data menjadi tidak *relevan* adalah salah dalam engetikan. Data yang tidak relevan juga lebih baik dihapus. Data pada pembersihan data yang dikelola jika berkurang

jumlah dan kompleksitasnya datanya maka akan mempengaruhi kinerja teknik data mining.

2) Penggabungan data kedalam database baru merupakan Integrasi data (data *integration*). Tidak hanya berasal dari satu database saja, namun dari beberapa *database* atau *file* teks yang dibutuhkan oleh data mining. Pada atribut yang mengidentifikasi *entitas* unik seperti atribut, nama, jenis produk, nomor pelanggan, dan lain-lain dilakukan integrasi data. Kesalahan integrasi data dapat menghasilkan hasil yang bias dan bahkan menyesatkan tindakan selanjutnya oleh karena itu pada integrasi data harus dilakukan dengan hati-hati. Contoh, produk menghasilkan penggabungan produk dari kategori yang berbeda, dengan integrasi data berdasarkan jenis maka tidak ada *korelasi* antara produk yang tidak ada.

3) Seleksi Data (Data *Selection*) Seringkali tidak semua data dalam database digunakan, sehingga hanya data yang dapat dianalisis yang diambil dari database *ID klien* saja.

4) Data digabungkan atau diubah dalam *format* yang sesuai untuk diproses dalam penambangan data (Data *Transformation*). *Format* data khusus sebelum dapat diterapkan memerlukan beberapa algoritma data mining. Contoh, analisis asosiasi dan pengelompokan hanya dapat menerima entri data kategorikal dari beberapa algoritma. Proses yang berkelanjutan harus dibagi menjadi beberapa *interval*. Proses ini sering disebut sebagai *transformasi* data.

5) Proses penting ketika algoritma diterapkan untuk menemukan pengetahuan yang berharga dan tersembunyi dari data adalah proses *mining*.

6) Evaluasi pola (*pattern evaluation*), Pada *fase* ini, hasil dari teknik data mining dievaluasi dalam bentuk model *tipikal* dan model *prediktif* untuk menilai apakah *hipotesis* yang ada benar-benar terwujud, ada beberapa *alternatif* yang dapat diambil, Jika ternyata hasil yang didapat tidak sesuai dengan *hipotesis*, seperti mencoba algoritma *mining* lain untuk data yang lebih sesuai atau menerima hasil ini sebagai hasil yang tidak diharapkan atau memberikan umpan balik untuk memperbaiki proses data *mining*.

7) *Visualisasi* dan penyajian pengetahuan tentang algoritma yang digunakan untuk memperoleh pengetahuan yang diperoleh oleh pengguna merupakan presentasi pengetahuan *(knowledge presentation)*. Dalam proses data *mining* adalah mengetahui bagaimana merumuskan keputusan atau tindakan berdasarkan hasil analisis yang dihasilkan ini merupakan langkah terakhir. Langkah yang diperlukan dalam proses data *mining* adalah penyajian hasil data *mining* sebagai pengetahuan yang mudah dipahami oleh siapa saja. *Visualisasi* juga dapat membantu mengkomunikasikan hasil data dalam presentasi.

2.4.1. Proses Data *Mining*

Proses yang biasanya dilakukan oleh data mining meliputi: deskripsi, prediksi, estimasi, klasifikasi, pengelompokan, dan asosiasi. Larose, 2005 proses data *mining* dijelaskan secara rinci sebagai berikut:

a) Deskripsi

Mengidentifikasi pola yang muncul berulang kali dalam suatu data dan mengubahnya menjadi aturan dan kriteria yang mudah dipahami oleh para ahli di bidang aplikasi merupakan tujuan deskripsi. Agar dapat meningkatkan pengetahuan sistem secara efektif Aturan yang dihasilkan harus bisa dipahami. Aktivitas data mining yang sering dibutuhkan dalam teknik *post-processing* untuk memvalidasi dan menjelaskan hasil dari proses data mining merupakan aktivitas deskriptif. Proses yang digunakan untuk memastikan bahwa hanya hasil yang *valid* dan bermanfaat yang dapat digunakan oleh pihak yang berkepentingan pada pasca pemrosesan.

b) Prediksi

Prediksi dan pemeringkatan memiliki kesamaan, berdasarkan perilaku atau nilai yang diharapkan di masa mendatang data dapat diurutkan.

c) Estimasi

Perkiraan variabel target lebih numerik daripada *kategoris* merupakan pengertian daripada estimasi. *Record* lengkap yang memberikan nilai variabel target sebagai nilai yang diharapkan menggunakan model yang dibangun. Berdasarkan nilai variabel *prediktor* nilai estimasi variabel target ditetapkan. Contoh, berdasarkan usia pasien, jenis kelamin, berat badan, dan kadar natrium darah, tekanan darah sistolik telah diperkirakan pada pasien rawat inap. Proses pembelajaran memunculkan model estimasi hubungan antara tekanan darah *sistolik* dengan nilai variabel *prediktor*.

d) Klasifikasi

Proses menemukan pola atau fungsi yang menggambarkan dan membedakan data ke dalam kelas-kelas merupakan klasifikasi. Proses memeriksa karakteristik objek dan menetapkan objek ke salah satu kelas yang telah ditentukan adalah keterlibatan klasifikasi.

e) Clustering

Pengelompokan data dalam kelas objek yang sama tanpa didasarkan pada kelas data tertentu adalah pengertian *clustering*. Kumpulan dari *record-record* yang memiliki kemiripan satu sama lain dan memiliki perbedaan dengan *record-record* dari cluster lainnya disebut juga *Cluster*. Menghasilkan pengelompokan objek yang terlihat sama dalam kelompok merupakan tujuanya. Semakin baik kualitas analisis *cluster* maka semakin besar perbedaan pada setiap *cluster* dan semakin besar kesamaan objek dalam suatu cluster.

2.5. Data *Science*

Menurut Pierre Simon Laplace dan Thomas Bayes Statistika merupakan dasar dari ilmu data. Ada empat cara, berbeda ilmu data dalam mengeksplorasi data, yaitu:

1) Pencarian Kenyataan

Data dapat dipulihkan dengan algoritma *pasif* atau aktif. Data dapat direpresentasikan sebagai *respons* dunia terhadap tindakan kita pada kebanyakan kasus data. Jika kita menganalisisnya, itu bisa sangat berharga saat membuat keputusan tindakan.

2) Pencarian Pola

Pendekatan *heuristik* lama yang bisa digunakan untuk pemecahan masalah, tetapi tidak semua masalah dapat diimplementasikan.

3) Memprediksi masa depan

Mampu memprediksi apa yang akan terjadi di masa depan dengan menggunakan data sampel adalah salah satu hal penting dalam *sains*. Keputusan untuk bertindak di masa depan. Tentu saja, dapat diprediksi di masa depan merupakan analisis *prediktif*.

4) Mengerti pengguna dan dunia.

Pemerintah serta perusahaan besar menyadari potensi ini, maka dari itu banyak yang berinvestasi besar-besaran dalam disiplin ilmu ini, seperti visi komputer, psikologi, dan ilmu saraf (Laura Igual & Santi Segui, 2017).

2.6. Pengertian Algoritma

Seperangkat prosedur yang diartikan dengan baik pada pengambilanl satu nilai, atau kumpulan nilai, sebagai *input* dan menghasilkan satu nilai atau kumpulan nilai sebagai *output* disebut Algoritma. Langkah-langkah perhitungan atau urutan yang dapat mengubah input menjadi *output* adalah algoritma (Thomas Cormen dan dkk, 2009).

2.6.1. Algoritma *Naive Bayes*

Algoritma klasifikasi yang menggunakan algoritma probabilistik dan statistik atau disebut *Naive Bayes* adalah algoritma yang diusulkan oleh ilmuwan

Inggris Thomas Bayes. Algoritma ini dikenal sebagai *teorema Bayes* yang artinya memprediksi peluang masa depan berdasarkan pengalaman masa lalu, sehingga kemandirian yang kuat *(naive)* dari kondisi/peristiwa apapun.

Naive Bayes dalam mengklasifikasi bekerja sangat baik dibandingkan dengan model pengklasifikasi lainnya, hal ini dibuktikan oleh jurnal yang berjudul "*Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages"* (Xhemali, Daniela, Chris J, Hinde dan Roger G. Stone, 2009). Klasifikasi *Naive Bayes* memiliki akurasi yang lebih baik daripada model pengklasifikasi lainnya. Kelebihan menggunakan metode ini adalah hanya membutuhkan sedikit data latih untuk menentukan estimasi parameter yang dibutuhkan untuk proses klasifikasi. Karena diasumsikan sebagai variabel *independen*, hanya *varians* variabel dalam kelas yang diperlukan untuk menentukan peringkat, bukan seluruh matriks *kovarians.*

Asumsi dari *Naive Bayes* adalah bahwa efek satu pada kelas tertentu tidak tergantung pada nilai atribut lainnya. Asumsi ini biasa disebut sebagai *independensi* kelas bersyarat. Itu dirancang untuk menyederhanakan perhitungan yang terlibat dan dalam hal ini disebut "*naive*". (Rezdy Anugrah Perdana hal.7). Pengklasifikasi probabilitas sederhana yang menerapkan *teorema Bayes* dengan asumsi *independensi* tinggi disebut *Naive Bayes Classifier* (NBC). Dengan demikian, label kelas catatan pengujian tidak dapat diprediksi dengan beberapa peristiwa meskipun set atributnya memilikki kesamaan pada beberapa contoh pelatihan. Situasi ini diperburuk oleh data yang bising atau adanya faktor pengganggu tertentu yang mempengaruhi klasifikasi tetapi tidak termasuk dalam

analisis. Algoritma berdasarkan probabilitas dan *teorema Bayesian* dengan asumsi bahwa setiap variabel adalah independen *(Independence)* dan mengasumsikan bahwa keberadaan suatu karakteristik (variabel) tidak ada hubungannya dengan keberadaan karakteristik (variabel) lainnya merupakan klasifikasi *Naive Bayes*. Model yang disederhanakan dari algoritma *Bayes*. *Nave Bayes* adalah apa yang digunakan dalam pembelajaran mesin sebagai algoritma untuk mendapatkan perkiraan suatu keputusan (Basuki,2006). Adapun Penerapan *teorema Bayes* dalam klasifikasi *Naive Bayes* sebagai berikut:

$$(Hi|E) = (E|Hi) * ((Hi)) \Sigma(E|Hk) * (p(Hk)) nk = 1$$

Keterangan :

P (Hi|E) = Probabilitas hipotesis Hi benar jika diberikan *evidence* (fakta) E.

P (E|Hi) = Probabilitas munculnya *evidence* (fakta) E jika diketahui Hipotesis Hi benar

P (Hi) = Probabilitas hipotesis Hi (menurut hasil sebelumnya) tanpa memandang *evidence* (fakta) apapun.

n = Jumlah hipotesis yang mungkin

Naïve Bayes Classifier umumnya memiliki karakteristik sebagai berikut :

a. Kuat untuk titik kebisingan yang terisolasi seperti titik rata-rata saat memperkirakan data probabilitas bersyarat. Pengklasifikasi *Naïve Bayes* dapat menangani nilai yang hilang dengan melewatkan sampel selama pemodelan dan klasifikasi.

b. Kelas probabilitas bersyarat untuk Xi tidak berdampak pada keseluruhan perhitungan probabilitas *posterior*, jika pada Xi adalah atribut yang tidak *relevan* maka kuat untuk atribut yang tidak *relevan*, , maka p (Xi ׀ Y) menjadi hampir terdistribusi secara merata..

c. Atribut terkait dapat menurunkan kinerja pengklasifikasi *Naïve Bayes* karena hipotesis kondisional independen tidak lagi mendukung atribut ini.

2.6.2. Algoritma *K-Nearest Neighbor*

Algoritma yang paling sederhana dalam pembelajaran mesin adalah algoritma *K-Nearest Neighbor*. Membuat model *machine learning* untuk menyimpan dataset pelatihan dan membuat prediksi untuk titik data baru, kemudian akan menemukan titik data terdekat merupakan cara penggunaanya. (Andreas C. Muller & Sarah Guido, 2017). Algoritma *K-Nearest Neighbor* merupakan Algoritma *supervised learning*. (Mustakim, Giantika Oktaviani, 2016).

Menggunakan algoritma *non parametrik* yang biasa digunakan dalam klasifikasi dan *regresi* merupakan prinsip dari algoritma *K-Nearest Neighbor* (Widodo Budiharto, 2016). Adapun langkah-langkah yang digunakan untuk adalah sebagai berikut:

1. Kumpulkan data dengan berbagai Algoritma.
2. Menghitung nilai jarak (*distance calculation*).

   Rumus *euclidian distance* dalam perhitungan jarak yang digunakan dapat dirumuskan sebagai berikut:

$$d\ (x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Jika, x = (x_1,....xn) dan y = (y_1,....yn), maka dapat dirumuskan sebagai berikut (Miroslav Kubat,2017):

$$d_E(x, y) = \sqrt{\sum_{1=1}^{n} (X_1 - X_2)^2}$$

3. Menganalisis data dengan berbagai algoritma.
4. Memproses data *training*
5. Menghitung *error rate*.
6. Jalankan algoritma untuk menentukan kelas mana yang cocok dengan data padamasukan data (Widodo Budiharto, 2016).

2.7. *Cross Validation*

Ada 3 (tiga) macam *Cross Validation*, yaitu:

1) Random Subsampling

Algoritma *cross validation* atau *Random subsampling.* Data latih dan data uji yang didistribusikan secara acak menggunakan algoritma ini. Data latih dan data uji pada kumpulan data yang sama dipilih secara acak dapat dilakukan beberapa kali untuk menghasilkan kesalahan estimasi rata-rata.

2) K-Fold Cross Validation

Data dibagi menjadi k bagian yang masing-masing memiliki jumlah data yang sama dalam teknik ini. Data uji digunakan sebagai bagian pertama dan sisanya digunakan sebagai data latih. Data uji baru, sisanya untuk data

latih baru dan seterusnya hingga maksimal k digunakan sebagai bagian kedua.

3) *Leave One Out Cross Validation*

Leave one out cross adalah turunan dari algoritma validasi silang *K Fold*, di mana *K* adalah konstanta yang sama dengan jumlah data. (Wesley, 2019).

2.8. Analisis Korelasi

Mengukur hubungan antara variabel X dan variabel Y dan sebaliknya merupakan cara untuk *koefisien* korelasi. Nilai korelasi variabel dimana 1 berarti variabel tersebut memiliki anti korelasi, nilai korelasi berkisar antara 1 sampai 1, artinya jika dikatakan variabel independen maka nilai korelasi 0 berarti tidak memiliki korelasi dan korelasi yang sangat kuat berarti 1. Ada dua jenis algoritma untuk mencari korelasi variabel, yaitu:

1) *Pearson Correlation Coefficient*

Koefisien korelasi *Pearson* menggunakan korelasi linier untuk mencari korelasi yang terus menerus bervariasi antara dua variabel. X dan Y dapat dianggap sebagai dua variabel yang memiliki n banyak sampel penelitian. Maka dapat dirumuskan sebagai berikut:

$$r = \frac{(N \sum X_i Y_i - \sum X_i \sum Y_i)}{\sqrt{NX_i^2 - (\sum X1)^2} \qquad \sqrt{NY_i^2 - (\sum Y_1)^2}}$$

Nilai r maksimum adalah 1 sedangkan nilai r terendah adalah 1. Jika nilai r adalah 1 maka X dan Y memiliki korelasi yang sangat kuat atau positif.

Jika nilai r adalah 0 maka X dan Y memiliki korelasi yang tidak jelas. Jika korelasi negatif, nilai r adalah 1 maka X dan Y tidak memiliki korelasi (Hong Hui Xu & Yong Deng, 2017).

1) Spearman Correlation Coefficient

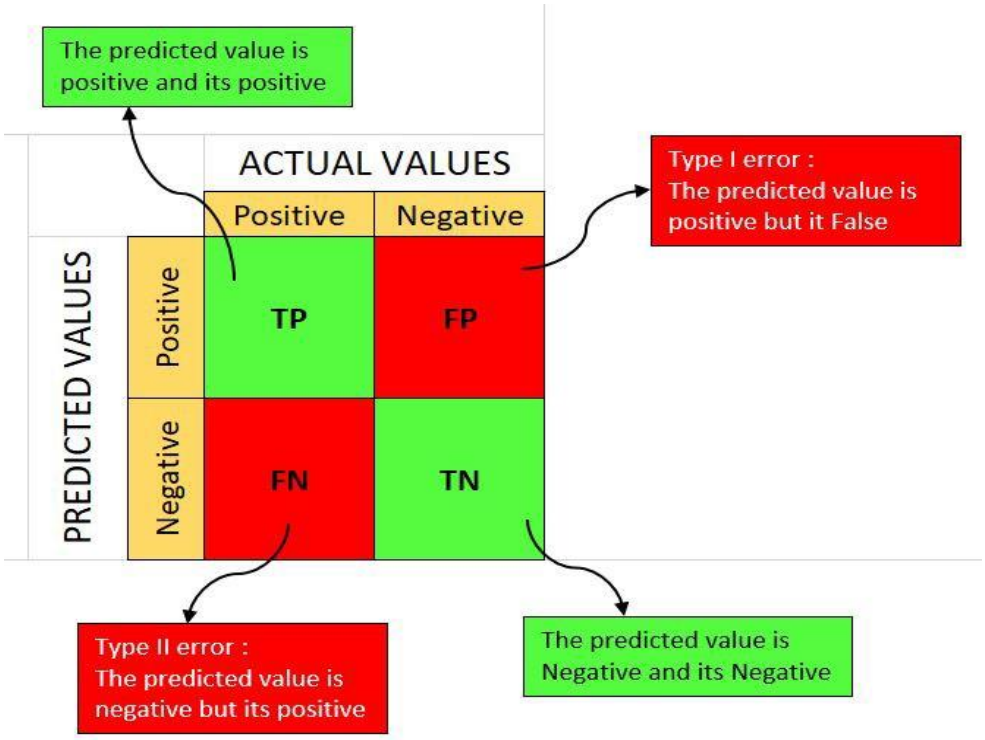
Korelasi *Spearman,* menggunakan sistem klasifikasi dan pada dasarnya untuk menghitung korelasi variabel data yang terdapat dalam variabel yang korelasinya ingin kita periksa, diasumsikan bahwa ada data (X_1,Y_1) dan (X_2,Y_2) dimana X_1 , X_2 dan Y_1 , Y_2 akan menghasilkan nilai positif yang memiliki korelasi yang cukup baik. Y_2 , Y_1 maka mendapatkan nilai negatif yang menunjukkan anti korelasi dari variabel yang diteliti. Rumus berikut digunakan dalam mencari korelasi variabel:

$$\rho = 1 - \frac{6 \sum d^2_1}{(n^2 - 1)}$$

Pada persamaan di atas dimana di adalah = (xi) - (yi) Ketika mencari nilai tertinggi dalam korelasi nonlinier, korelasi Spearman kurang sensitif dibandingkan dengan korelasi Pearson (Steven S. Skiena, 2018)

2.9. *Confusio*n Matriks

Matriks konfusi adalah ringkasan tabel dari jumlah prediksi yang benar dan salah yang dibuat oleh pengklasifikasi, biasanya digunakan untuk mengukur kinerja model klasifikasi dan juga dapat digunakan untuk mengevaluasi kinerja model klasifikasi melalui perhitungan metrik kinerja seperti akurasi, presisi, perolehan, dan skor F1.



Gambar 2.2. Matriks Konfusi

Dimana :

1. True Positive (TP) memprediksi ramalan positif dan itu benar.
2. True Negative (TN) memprediksi prediksi negatif dan itu benar.
3. False Positive (FP) : (Kesalahan Tipe 1) adalah memprediksi positif dan itu salah.
4. False Negative (FN) : (Kesalahan tipe 2 ini sangat berbahaya) adalah untuk memprediksi negatif dan salah.

2.9.1. Akurasi

Identifikasi atau klasifikasi diharapkan dapat mengklasifikasikan semua kumpulan data dengan benar melakukan dilakukan sistem. Namun tidak dapat dipungkiri bahwa kinerja suatu sistem yang melakukan klasifikasi tidak dapat selalu bisa 100% benar. Oleh karena itu, sistem harus diukur kinerjanya. Secara

umum matriks konfusi digunakan dalam mengukur kinerja klasifikasi (Prasetyo, 2014). Penting untuk mengukur kinerja sistem penilaian. Kemampuan sistem untuk mengklasifikasikan data menggambarkan kinerja sistem klasifikasi. Salah satu algoritma yang dapat digunakan untuk mengukur performansi suatu algoritma klasifikasi adalah *Confusion matrix*. Informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi yang seharusnya, merupakan isi dari *confusion matrix*.

Terdapat 4 (empat) istilah yang mewakili hasil dari proses klasifikasi pada pengukuran kinerja dengan menggunakan *confusion matrix*. Pada keempat diantaranya, *True Positive* (TP), *True Negative* (TN), *False Positive* (FP) dan *False Negative* (FN). Data negatif tetapi dilaporkan sebagai data positif adalah *False Positive* (FP), jumlah data negatif yang terdeteksi dengan benar adalah *True Negative* (TN), data positif yang terdeteksi dengan benar adalah *true positives* (TP). Negatif palsu (FN) adalah kebalikan dari positif sejati, sehingga datanya positif, tetapi terdeteksi sebagai data negatif. Pada tipe klasifikasi biner yang hanya memiliki 2 keluaran kelas, matriks konfusi dapat disajikan seperti pada Tabel 2.2 sebagai berikut:

Tabel 2.2. *Confusion Matrix* Klasifikasi Biner

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

Untuk mendapatkan nilai akurasi maka nilai *True Negative* (TN), *False Positive* (FP), *False Negative* (FN) dan *True Positive* (TP). Seberapa tepat sistem pada pengklasifikasian data dengan benar digambarkan oleh nilai akurasi. Dengan kata lain, nilai presisi adalah prediksi antara data yang diklasifikasikan dengan benar dan kumpulan data. Nilai akurasi dapat diperoleh dengan rumus:

$$\textbf{Akurasi} = \frac{\text{Jumlah data yang diprediksi secara benar}}{\text{Jumlah pengujian yang dilakukan}} \text{ x } 100\%$$