BAB II
LANDASAN TEORI

2.1. Penelitian Terdahulu

Penelitian mengenai analisis klasifikasi menurut Rafly Dikhi Firdaus dengan judul Penerapan Data Mining untuk menentukan Klasifikasi Kinerja Pegawai Pada Sistem Informasi Kepagawaian, tahun terbit 2017. Dengan penilaian prestasi kerja yang objektif akan memberikan umpan balik dilakukan secara objektif dan menyeluruh. Bawahan lambat laun akan memahami objektifitas kerja dan mampu mendorong iklim produktivitas perusahaan menggunakan metode data mining klasifikasi *Naïve Bayes* dan mendapatkan hasil berdasarkan data pegawai yang diuji menggunakan SIK yang dijadikan data *training*, metode *Naïve Bayes* berhasil mengklasifikasikan 49 data dari 50 data yang diuji. Sehingga dengan demikian metode *Naïve Bayes* ini berhasil memprediksi kinerja karyawan dengan persentase keakuratan sebesar 98%. [7]

Persamaan penelitian ini dengan penelitian yang dilakukan oleh Rafly Firdaus adalah penelitian sama-sama melakukan analisis menggunakan metode *Naïve Bayes* berdasarkan data pada sistem. Perbedaannya permasalahan yang terjadi dalam persusahan adalah kesulitan dalam mengambil penilaian prestasi kerja bawahan yang objektif sehingga dalam penelitian ini dilakukan klasifikasi prestasi kerja yang mampu mendorong iklim produktivitas perusahaan. Sedangkan di dalam perguruan tinggi UNIKOM dalam penilaian dosen dengan

menyeluruh program studi yang belum objektif dalam perguruan tinggi sehingga dilakukan klasifikasi opini program studi yang mampu mendorong produktivitas.

Penelitian terdahulu kedua menurut Mujib Ridwan pada tahun 2013 dengan judul Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa menggunakan Algoritma *Naïve Bayes Classifier*. Penelitian ini mengevaluasi kinerja akademik mahasiswa pada tahun ke-2 dan diklasifikasikan dalam kategori mahasiswa yang dapat lulus tepat waktu atau tidak. Sampel yang dimabil data mahasiswa angkatan 2005-2009 yang sudah dinyatakan lulus akan digunakan sebagai data target. Menggunakan metode data mining algoritma *Naïve Bayes Classifier (NBC)* untuk membentuk *table* probabilitas sebagai dasar proses klasifikasi kelulusan mahasiswa. Hasil klasifikasi kinerja akademik mahasiswa menunjukkan bahwa faktor yang paling berpengaruh dalam penentuan klasifikasi kinerja akademik mahasiswa yaitu Indeks Prestasi Kumulatif (IPK), Indeks Prestasi (IP) semester 1, IP semester 4, dan jenis kelamin. Pengujian pada data mahasiswa angkatan 2005-2009, algoritma NBC menghasilkan nilai *precision* 83%, *recall* 50% dan *accurary* 70%.[4]

Penelitian terdahulu mengenai analisis klasifikasi selanjutnya menurut Indah Purnamasari, Karnita Afnisari yang berjudul Performansi Klasifikasi Dosen Berprestasi menggunakan metode *Naïve Bayes Classifier*. Pendidik yang berprestasi harus diberi penghargaan yang sesuai. Penelitian ini bertujuan untuk memotivasi para pendidik agar menumbuhkan dedikasi yang tinggi terhadap

terwujudnya peserta didik yang cerdas dan menumbuhkan rasa bangga terhadap profesinya.

Pendidik di perguruan tinggi disebut dosen. Prestasi dosen berprestasi adalah dosen pelaksana Tridharma. Perguruan Tinggi yaitu Pendidikan, Penelitian dan Pelayanan kepada masyarakat. Namun pemilihan dosen berprestasi sesuai dengan persyaratan sistem penghargaan yang ditetapkan pemerintah tentu bukan hal yang mudah. Oleh karena itu, untuk membantu pemilihan dosen berprestasi dalam penelitian ini digunakan klasifikasi data mining dengan metode dari *Naive Bayes Classifier* dengan hasil penelitian ini mencapai akurasi sebesar 91,67%.[5]

Penelitian terdahulu keempat menurut Viny Novika Sari, Lola Yorita Astri, dan Errissya Rasywir yang berjudul Analisis Dan Penerapan Algoritma *Naive Bayes* Untuk Evaluasi Kinerja Karyawan Pada PT. Pelita Wira Sejahtera. Pada PT. Pelita Wira Sejahtera penilaian terkadang dilakukan secara subjektif dan keterbatasan dalam mengontrol setiap karyawan yang berkerja. Maka penulis melakukan analisis data mining pada data-data penilaian karyawan untuk mengetahui mana karyawan yang memiliki kinerja yang sangat baik, baik, cukup, kurang. Data yang digunakan sebanyak 149 data yang kemudian disajikan kedalam format arff. Metode yang digunakan klasifikasi *Naïve Bayes* dengan persentasi akurasi terbesar diperoleh dengan menggunakan *Use Training Set Correctly* yaitu sebesar 95,302%, menggunakan *5-Fold Cross Validation Correctly* sebesar 93,9597% dan menggunakan *10-Fold CrossValidation* sebesar 93,9597%. Hasil seleksi atribut menggunakan algoritma *classifier attribute*

evaluation dinyatakan bahwa atribut yang paling berpengaruh terhadap klasifikasi penilaian kinerja adalah orientasi pada efisiensi.[6]

Penelitian terdahulu kelima menurut M.Syukri Mustafa, M.Rizky Ramadhan, Angelina yang berjudul Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma *Naïve Bayes Classifier*. Penelitian ini mengevaluasi kinerja akademik mahasiswa STMIK Diponegoro Makasar pada 2 tahun pertama dengan menggunakan teknik data mining algoritma *Naïve Bayes Classifier (NBC)* untuk membentuk tabel probabilitas sebagai dasar proses klasifikasi kinerja akademik mahasiswa yang kelulusannya akan diklasifikasikan dan memberikan rekomendasi untuk proses kelulusan tepat waktu yang paling tepat dengan nilai optimal. Hasil yang diperoleh dari penelitian ini menunjukkan bahwa faktor yang paling mempengaruhi penentuan klasifikasi kinerja akademik mahasiswa adalah Indeks Prestasi (IP) pada semeseter 1,2,3,4 dan jenis kelamin. Pengujian pada beberapa data mahasiswa angkatan 2008-2011 yang diambil secara acak, algoritma *NBC* menghasilkan nilai akurasi 92,3%. [8]

Penjelasan penelitian terdahulu di atas mengenai evaluasi klasifikasi kinerja, dan penelitian kasus yang berbeda dengan metode yang sama. Penelitian terdahulu yang telah dipaparkan akan diuraikan secara singkat pada Tabel 2.1

Tabel 2.1 Ringkasan penelitian terdahulu

No	Peneliti	Judul Penelitian	Metode Penelitian	Hasil Penelitian
1.	Rafly Dikhi Firdaus	Penerapan Data Mining untuk menentukan Klasifikasi Kinerja Pegawai Pada Sistem Informasi Kepagawaian	*Naïve Bayes*	Bawahan lambat laun akan memahami objektifitas kerja dan mampu mendorong iklim produktivitas perusahaan menggunakan metode data mining klasifikasi *Naïve Bayes* dan mendapatkan hasil berdasarkan data pegawai yang diuji menggunakan SIK yang dijadikan data *training*, metode *Naïve Bayes* berhasil mengklasifikasikan 49 data dari 50 data yang diuji. Sehingga dengan demikian metode *Naïve Bayes* ini berhasil memprediksi kinerja

No	Peneliti	Judul Penelitian	Metode Penelitian	Hasil Penelitian
				karyawan dengan persentase keakuratan sebesar 98%.
2.	Mujib Ridwan	Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa menggunakan Algoritma *Naïve Bayes Classifier*	*Naïve Bayes Classifier (NBC)*	Hasil klasifikasi kinerja akademik mahasiswa menunjukkan bahwa faktor yang paling berpengaruh dalam penentuan klasifikasi kinerja akademik mahasiswa yaitu Indeks Prestasi Komulatif (IPK), Indeks Prestasi (IP) semester 1, IP semester 4, dan jenis kelamin. Pengujian pada data mahasiswa angkatan 2005-2009, algoritma NBC menghasilkan nilai *precision* 83%, *recall* 50% dan *accurary* 70%.

No	Peneliti	Judul Penelitian	Metode Penelitian	Hasil Penelitian
3.	Indah Purnamasari, Karnita Afnisari	Performansi Klasifikasi Dosen Berprestasi menggunakan metode *Naïve Bayes Classifier*	*Naïve Bayes Classifier*	Pendidik di perguruan tinggi disebut dosen. Prestasi dosen berprestasi adalah dosen pelaksana Tridharma. Perguruan Tinggi yaitu Pendidikan, Penelitian dan Pelayanan kepada masyarakat. Namun pemilihan dosen berprestasi sesuai dengan persyaratan sistem penghargaan yang ditetapkan pemerintah tentu bukan hal yang mudah. Oleh karena itu, untuk membantu pemilihan dosen berprestasi dalam penelitian ini digunakan klasifikasi data mining dengan metode dari *Naive Bayes Classifier* dengan hasil penelitian ini

No	Peneliti	Judul Penelitian	Metode Penelitian	Hasil Penelitian
				mencapai akurasi sebesar 91,67%
4.	M.Syukri Mustafa, M.Rizky Ramadhan, Angelina	Implementasi Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma *Naïve Bayes Classifier*	*Naïve Bayes Classifier (NBC)*	Hasil yang diperoleh dari penelitian ini menunjukkan bahwa faktor yang paling mempengaruhi penentuan klasifikasi kinerja akademik mahasiswa adalah Indeks Prestasi (IP) pada semeseter 1,2,3,4 dan jenis kelamin. Pengujian pada beberapa data mahasiswa angkatan 2008-2011 yang diambil secara acak, algoritma *NBC* menghasilkan nilai akurasi 92,3%.

Berikut ini adalah perbedaan penelitian yang penulis teliti dengan penelitian sebelumnya:

a) Data yang digunakan oleh penulis dalam penelitian ini adalah data evaluasi program studi UNIKOM semester ganjil tahun 2017/2018, semester genap tahun 2017/2018, semester ganjil tahun 2018/2019, semester genap 2018/2019, semester ganjil 2019/2020, semester genap 2019/2020, dan semester ganjil 2020/2021.
b) Data yang dianalisa dan diproses penulis akan dijadikan untuk klasifikasi program studi UNIKOM dengan parameter berupa data pengajaran dan mengajar.
c) Metode yang digunakan dalam penelitian ini oleh penulis yaitu algoritma *Naïve Bayes Classifier*, yang digunakan untuk mengklasifikasi program studi UNIKOM. Untuk melihat tingkat keakurasian data yang diteliti menggunakan *tools jupyter notebook python* untuk mengolah data dalam klasifikasi evaluasi program studi UNIKOM.

2.2. Data Mining

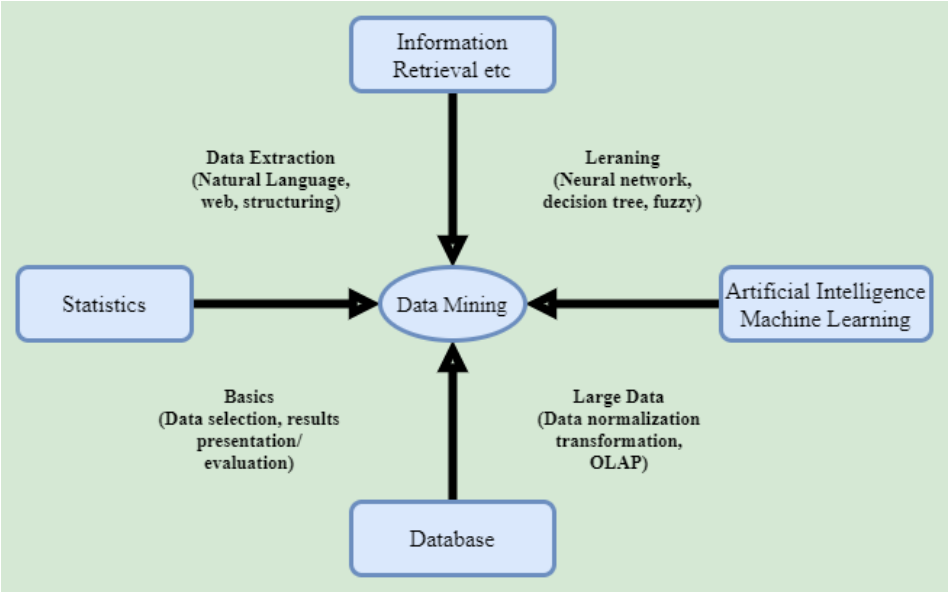
2.2.1. Pengertian Data Mining

Menurut David Hand, Heikki Mannila, dan Padhraic Smyth dari *Massachusetts Institute of Technology*, data mining adalah analisis data (*big data*) untuk menemukan hubungan yang jelas, dan untuk menarik kesimpulan dengan cara yang saat ini berguna untuk data dan pemilik kesimpulan yang sebelumnya tidak diketahui.

Menurut Yuli Mardi, data mining adalah proses menggunakan teknik atau metode tertentu untuk menemukan pola atau informasi yang menarik dalam data yang dipilih. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Memilih metode atau algoritma yang tepat sebenarnya tergantung pada tujuan dan proses *Knowledge Discovery in Database (KDD*) secara keseluruhan dalam *database*. [9]

Data mining adalah proses menggunakan statistik, matematika, kecerdasan buatan dan teknik pembelajaran mesin untuk mengekstrak dan mengidentifikasi informasi yang berguna dan pengetahuan terkait dari database besar. Dilihat dari definisi yang disampaikan, hal-hal penting terkait data mining [13]:

1. Data mining adalah proses otomatis dari data yang ada.
2. Data yang akan diolah adalah data yang sangat besar.
3. Tujuan dari data mining adalah untuk menemukan hubungan atau pola yang dapat memberikan indikasi yang berguna.



Gambar 2.1 Bidang Ilmu Data Mining

(Sumber: Mardi, Yuli.Edik Informatika, 2017)

Salah satu kesulitan dalam mendefinisikan data mining adalah bahwa data mining mewarisi banyak aspek dan teknik dari disiplin ilmu yang telah ditetapkan sebelumnya. Gambar 2.1 menunjukkan bahwa data mining berasal dari berbagai disiplin ilmu, seperti kecerdasan buatan, *machine learning*, statistik, *database*, dan pencarian informasi. [9]

2.2.2. Metode Data Mining

Data mining memiliki banyak metode atau fungsi yang dapat digunakan untuk menggali dan menemukan pengetahuan. Menurut Susanto & Suryadi, [8] data mining memiliki enam kelompok fungsional, yaitu:

1. Description (deskripsi) member gambaran tentang sejumlah data besar dan memiliki jenis data. Metode ini seperti, *Decision tree* (pohon keputusan), *Exploratory Data Analysis* (analisis data eksplorasi), dan *Neural Network* (metode jaringan saraf).

2. *Estimation* (Estimasi), menebak nilai yang tidak diketahui, seperti menebak pendapatan seseorang ketika mengetahui beberapa informasi tentang orang tersebut. Metode yang dapat digunakan adalah *Point estimation, Confidence Interval Estimations, Simple Linear Regression, Correlation,* dan *Multiple Regression.*
3. *Prediction* (Prediksi), memperkirakan nilai masa depan, seperti peramalan persediaan komoditas dalam tiga tahun ke depan. Fungsi ini meliputi *Neural Network, Decision tree,* dan *k-nearest neighbor*.
4. *Classification* (Klasifikasi) adalah proses mencari model atau fungsi yang dapat membedakan konsep atau kategori data.Tujuannya adalah untuk dapat mengestimasi kategori objek dengan label yang tidak diketahui. Fungsi ini meliputi *Neural Network, Decision Tree*, *k-Nearest Neighbor* dan *Naive Bayes*.
5. *Clustering* (Pengelompokan) adalah proses mengidentifikasi data dengan karakteristik tertentu. Fitur ini meliputi model *Hierarchical Clustering*, metode *K-Means*, dan *Self Organizing Map (SOM).*
6. Association *(*Asosiasi), juga dikenal sebagai analisis keranjang belanja, di mana fungsi ini digunakan untuk mengidentifikasi item produk yang mungkin dibeli konsumen dengan produk lain. Dalam metode atau algoritma yang termasuk dalam fungsi ini meliputi *Apriori, Generalized Seqeuntial Pattern (GSP), FP-Growth,* dan *GRI Algorithm.*

2.3. *Python*

Python adalah suatu bahasa pemrograman yang lumayan populer yang mempunyai banyak khasiat buat menunjang pemrograman yang berorientasi objek serta bisa berjalan diberbagai berbagai platform sistem pembedahan semacam *PCs, Macintosh, UNIX*. Sebagian kelebihan dari bahasa pemrograman *python* diantara lain:

a. Pengembangan program dicoba dengan kilat serta *coding* yang lebih sedikit;
b. Menunjang multi platform;
c. *Python* gampang dipelajari;
d. Mempunyai sistem pengelolaan memori yang otomatis;
e. *Python* bertabiat *Object Oriented Programming*.

Sejarah *Python*

Python dikembangkan oleh seseorang kebangsaan Belanda tepatnya berasal dari kota Amsterdam bernama Guido Van Rossum pada tahun 1990 yang artinya kelanjutan asal bahasa pemrograman ABC. Tahun 1995, Guido pindah ke CNRI dengan melanjutkan pengembangan *Python*. Versi terakhir yang dikeluarkan ialah 1.6. Tahun 2000, Guido serta para pengembang inti *Python* pindah ke BeOpen.com yang merupakan sebuah perusahaan komersial dan membentuk BeOpen Python Labs. *Python* 2.0 merupakan versi yang dikeluarkan oleh BeOpen, setelah mengeluarkan *python 2.0* Guido serta beberapa anggota tim *python labs* pindah ke *Digital Creations.* Saat ini pengembangan *Python* terus dilakukan oleh sekumpulan pemrograman yang

dikoordinir oleh Guido dan *python software foundation. Python Software Foundation* merupakan sebuah organisasi yang dibentuk sebagai pemegang hak cipta *Python* sejak adanya versi 2.1 dan dengan itu untuk mencegah Python dimiliki oleh perusahaan komersial. Pada saat ini distribusi *Python* sudah mencapai versi 2.6.1 dan versi 3.0. Nama *Python* dipilih oleh Guido sebagai nama bahasa ciptaanya karena Guido bangga dengan kecintaan acara televise *Monty Python's Flying Circus*, oleh sebab itu sering terjadi ungkapan khas dari acara tersebut muncul dalam korespondensi antar pengguna *Python.*

2.4. Library

1. Pandas

Pandas merupakan perlengkapan analisis serta manipulasi informasi sumber terbuka yang kilat, kokoh, fleksibel, serta gampang digunakan, dibentuk di atas bahasa pemrograman *Python*.

2. *Numpy*

NumPy (Numeric Python) adalah salah satu dasar penting untuk komputasi medis dengan *Python*. Ini adalah pustaka *Python* yang menawarkan objek *array* multi dimensi, objek turunan beragam (yang mencakup array dan matriks bertopeng), dan kumpulan latihan untuk operasi instan pada *array*, yang mencakup matematika, logika, manipulasi bentuk, pengurutan, pemilihan, *I/ O*, transformasi *Fourier* diskrit, aljabar linier sederhana, operasi statistik sederhana, simulasi acak, dan banyak lagi.

3. *NLTK*

NLTK (Naturel Language Toolkit) merupakan platform termuka dalam menciptakan program python untuk bekerja menggunakan data bahasa insan. Ini menyediakan anatar muka yang simpel di gunakan ke lebih dari 50 corpora dan *lexical resources* seperti WordNet, beserta menggunakan rangkaian perpustakaan pemrosesan teks buat klasifikasi, tokenisasi, *stemming, tagging, parsing* dan penalaran sematik, perpustakaan NLP kekuatan industri dan forum diskusi yang aktif.

4. *Sklearn*

Scikit-Learn atau *SKleran Machine Learning in Python* merupakan Alat sederhana dan efisien untuk penambangan data dan analisis data. Dapat diakses oleh semua orang, dan dapat digunakan kembali dalam berbagai konteks, dibangun di NumPy, SciPy, dan matplotlib, dan Sumber terbuka, dapat digunakan secara komersial - lisensi BSD.

Dalam *SKlearn* terdapat banyak modul seperti *classification, regression, clustering, preprocessing, model selection.* Metode *Support Vector Classification* dapat diperluas untuk memecahkan masalah regresi. Metode ini disebut *Support Vector Regression*. Model yang dihasilkan oleh klasifikasi vektor pendukung (seperti dijelaskan di atas) hanya bergantung pada subset dari data pelatihan karena fungsi biaya untuk membangun model tidak peduli dengan titik pelatihan yang berada di luar *margin*. Secara analog, model yang dihasilkan oleh *Support Vector Regression* hanya bergantung pada subset dari data pelatihan, karena

fungsi biaya untuk membangun model mengabaikan data pelatihan yang dekat dengan prediksi model.

5. *Seaborn*

Seaborn adalah pustaka visualisasi data Python berdasarkan matplotlib. Ini menyediakan antarmuka tingkat tinggi untuk menggambar grafik statistik yang menarik dan informatif.

2.5. Pengertian Klasifikasi

Klasifikasi berasal dari bahasa latin yaitu *classis* yang berarti mengelompokkan benda-benda yang sejenis dan memisahkan benda-benda yang berbeda; "*classify*" dalam bahasa Inggris adalah kumpulan bahan pustaka (buku, brosur, peta, kaset video, rekaman suara, dll) yang disusun menurut sistem klasifikasinya berbasis fitur-fitur (aspek) masing-masing bahan pustaka.

Sedangkan klasifikasi menurut terminologi adalah proses membagi objek atau konsep secara logis ke dalam kelas hierarkis, subkelas, dan subkelas berdasarkan titik objek atau konsep yang sama dan berbeda. Secara umum klasifikasi juga diartikan sebagai penataan pengetahuan umum ke dalam sejumlah kegiatan yang disusun secara sistematis. (Hasby, 2012: 40)

Pengklasifikasian adalah proses pengelompokan objek ke dalam kategori atau kategori yang telah ditentukan. Klasifikasi terjadi dalam berbagai aktivitas manusia, salah satunya adalah untuk menentukan kualitas produk. Apakah pasien dipengaruhi oleh penyakit yang diamati atau tidak, mereka dapat diklasifikasikan dengan mengamati karakteristik ini.

Saat mengklasifikasikan variabel, analisis klasifikasi diperlukan. Analisis klasifikasi adalah metode untuk menganalisis hubungan antara beberapa variabel prediktor dan variabel respons sebagai variabel kualitatif. Beberapa prediktor tersebut akan digunakan untuk memprediksi kategori atau kategori dari variabel respon. Metode yang digunakan untuk klasifikasi kategori pertama adalah dengan memprediksi probabilitas variabel kualitatif pada setiap kategori sebagai dasar klasifikasi (James et al, 2013).

Manfaat Klasifikasi

Klasifikasi dalam beberapa hal digunakan sebagai prosedur data mining dan memiliki banyak manfaat dalam kehidupan sehari-hari. Adapun beberapa manfaat dari analisis klasifikasi adalah sebagai berikut:

a. Berguna dalam hal iklan dimana iklan dilakukan di tempat-tempat atau orang-orang tertentu yang mempunyai kecenderungan tertarik pada suatu produk yang diiklankan.
b. Dalam perbankan klasifikasi dapat digunakan untuk memutuskan apakan seseorang layak diberi pinjaman atau tidak.
c. Klasifikasi digunakan dalam mesin pencari di internet dimana akan dapat menyasar seseorang dengan kriteria tertentu.

2.6. Algoritma *Naïve Bayes Classifier*

2.6.1. Pengertian Algoritma

Algoritma adalah proses perhitungan yang mengambil nilai-nilai tertentu atau seperangkat nilai sebagai input dan kemudian diproses sebagai *output* sehingga algoritma adalah serangkaian langkah perhitungan yang mengubah *input output* (Comen, 2009:5).

Metode pengklasifikasian menggunakan probabilitas dan statistik yang pertama kali ditemukan oleh ilmuwan Inggris bernama *Thomas Bayes*, yaitu suatu metode memprediksi peluang di masa depan berdasarkan masa sebelumnya, dengan terkenal metode *Teorema Bayes*. Ciri khusus dari *Naïve Bayes Classifier* adalah asumsi yang sangat kuat akan independensi dari masing-masing kejadian atau kelas[8].

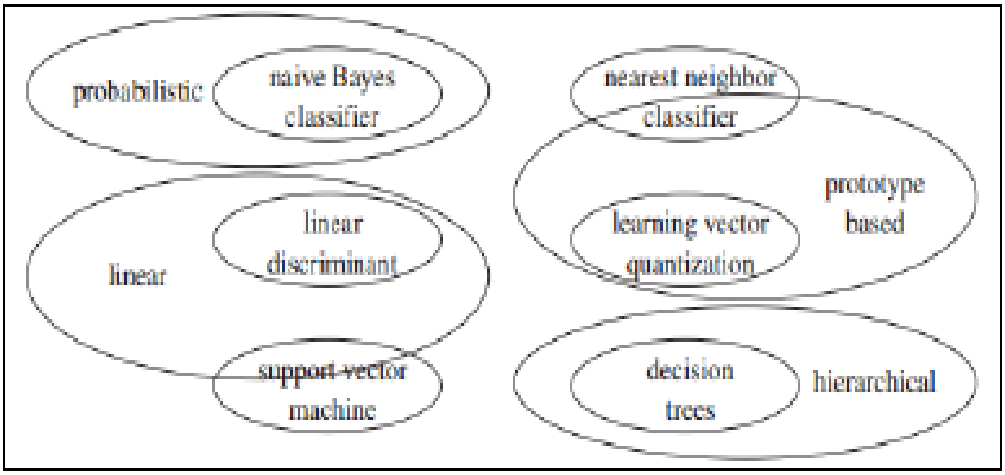
Menurut Olson Delen (2008), Algoritma menjelaskan *Naive Bayes* dari setiap kelas keputusan, dan menghitung probabilitas bahwa kelas keputusan itu benar mengingat vektor informasi objek. Algoritma mengasumsikan bahwa properti objek adalah *independen*. Probabilitas yang terlibat dalam menghasilkan perkiraan akhir dihitung sebagai jumlah frekuensi dalam tabel keputusan

2.6.2. Pengertian Naive Bayes Classifier

Salah satu metode data mining adalah klasifikasi data, yaitu memetakan data ke suatu kelas atau beberapa kelas yang sebelumnya sudah didefinisikan. Salah satu metode klasifikasi data adalah *Naive Bayes Classifier (NBC). Naïve Bayes*

Classifier adalah pengklasifikasi statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas.

Pada gambar dapat dilihat skema yang sering digunakan dalam proses metode klasifikasi dimana terdapat *Naïve Bayes Classifier[8]*.



Gambar 2.2 Skema Klasifikasi Algoritma NBC

(Sumber: Creative Information Technology Journal 4.2 (2018): 151-162.)

Klasifikasi *Naive Bayes* didasarkan pada *teorema Bayes* dan memiliki kemampuan klasifikasi yang mirip dengan pohon keputusan dan jaringan saraf. Klasifikasi *Naive Bayes* terbukti memiliki akurasi dan kecepatan yang tinggi ketika diterapkan pada *database big* (Kusrini, 2009).*Teorema bayes* memiliki bentuk umum sebagai berikut:

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)}$$

Persamaan (1)

Keterangan:

X = Data dengan kelas belum diketahui

H = Hipotesa data X merupakan suatu kelas spesifik

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X (*posteriori probability*)

P(H) = Pobabilitas hipotesis H *(prior probability*)

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

P(X) = Probabilitas dari X

Menurut buku dari Eko Prasetyo, ide pokok dari aturan *Bayes* yaitu dengan hasil dari hipotesis atau peristiwa (H) dapat diperkirakan berdasarkan bukti (X) yang diamati. Ada beberapa hal penting yang perlu diperhatikan yaitu:

[1] Sebuah probabilitas awal/*prior* H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.

[2] Sebuah probalilitas akhir H atau P(H|X) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

Penjabaran lebih lanjut rumus *Naïve Bayes* tersebut dilakukan dengan menjabarkan secara terperinci (C|X1…,Xn) menggunakan aturan perkalian sebagai berikut.

$P(C|x_1,.....,x_n = P(C)P(x_1,….,x_n \mid C)$

$= P(C)P(X_1|C)P(X_2,….,X_n|C,X_1)$

$= (C)P(X_1|C)P(X_2|C,X_1)P(X_3,...X_n|C,X_1,X_2(C)P(X_1|C)$

$P(X_2|C,X_1)P(X_3|C,X_1,X_2)P(X_4,....,X_n|C,X_1,X_2,X_3)P(C)$

$= P(X_1|C)P(X_2|C,X_1)P(X_3|C,X_1,X_2)....$

$P(X_n|C,X_1,X_2,X_3,.....X_{n-1}.....$ persamaan (2)

Jika semakin banyak factor yang kompleks mempengaruhi nilai probabilitas semakin tidak mungkin untuk menghitung nilai-nilai tersebut satu per satu. Proses perhitungan akan semakin sulit dilakukan, sehingga digunakan asumsi independensi yang sangat tinggi, yaitu setiap atribut dapat saling independen. Dengan asumsi tersebut diperlukan persamaan 3:

$$P(X_i|X_j) = \frac{P(X_i|X_j)}{P(X_j)} = \frac{P(X_i \cap X_j)}{P(X_i)} = P(X_i)$$

Untuk i≠j, sehingga

$$P(X_i|C,X_j = P(X_i|C) \quad \text{persamaan (3)}$$

Dari persamaan 2 tersebut dapat diambil kesimpulan bahwa asumsi independensi membuat syarat perhitungan menjadi lebih sederhana. Selanjutnya penjabaran (P(C|X1,…..,Xn) dapat disederhanakan menjadi persamaan 4 :

$$P(X2|C)P(X3|C) \dots P(C|X1, \dots Xn) = P(X1|C) = \prod_{i=1}^{n} p(Xi|C) \quad \text{persamaan (4)}$$

Keterangan:

$\prod_{i=1}^{n} p(Xi|C)$ = perkalian ranting antar atribut

Persamaan 4 merupakan *teorema Bayes*, yang selanjutnya akan digunakan untuk melakukan perhitungan klasifikasi. Untuk klasifikasi data kontinu atau *numerik*, menggunakan rumus distribusi Gaussian dengan 2 parameter:

Mean μ dan varian σ:

$$P(X_i = X_i|C = c_j) = \frac{1}{\sqrt{2\pi\sigma ij}} exp \frac{(xi-\mu ij)2}{2\sigma 2ij} \quad \text{persamaan (5)}$$

Keterangan:

P : Peluang

X_i : Atribut ke i

X_j : Nilai atribut ke i

C : Kelas yang dicari

Cj : Sub kelas Y yang dicari

μ : Menyatakan rata–rata dari seluruh atribut

σ : Deviasi standar, menyatakan varian dari seluruh atribut.

Teorema Naïve Bayes memiliki beberapa kelebihan dan kekurangan yaitu sebagai berikut.

Kelebihan:

Teori *Bayesian*, mempunyai beberapa kelebihan (Grainner 1998), yaitu:

a. Mudah untuk dipahami;
b. Hanya memerlukan pengkodean yang sederhana;
c. Lebih cepat dalam penghitungan;
d. Menangani kuantitatif dan data diskrit;
e. Kokoh untuk titik *noise* yang diisolasi, misalkan titik yang dirata – ratakan ketika mengestimasi peluang bersyarat data;
f. Hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata – rata dan variansi dari variabel) yang dibutuhkan untuk klasifikasi;

g. Menangani nilai yang hilang dengan mengabaikan instansi selama perhitungan estimasi peluang;
h. Cepat dan efisiensi ruang;
i. Kokoh terhadap atribut yang tidak relevan.

Kekurangan:

Sedangkan kekurangan dari *Teorema* Bayes adalah :

a. Kekurangan dari Teori probabilitas *Bayesian* yang banyak dikritisi oleh para ilmuwan adalah karena pada teori ini, satu probabilitas saja tidak bisa mengukur seberapa dalam tingkat keakuratannya. Dengan kata lain, kurang bukti untuk membuktikan kebenaran jawaban yang dihasilkan dari teori ini.
b. Tidak berlaku jika probabilitas kondisionalnya adalah 0 (nol), apabila nol maka probabilitas prediksi akan bernilai nol juga
c. Mengasumsikan variabel bebas

2.3.1. Metode Naïve Bayes Classifier

Adapun alur dilakukan proses dari metode *Naïve Bayes Classifier* sebagai berikut:

Gambar 2.3 Alur Metode *Naïve Bayes*

1. Input Data

   Input data untuk memasukkan data dari hasil pengumpulan data yang sudah berbentuk format csv (*Comma Separated Values*) ke program dalam *python*.

2. Pelabelan

   Pada pelabelan ini untuk menentukan variabel independen (varibael yang mempengaruhi terjadinya timbulnya varibael dependen/terikat) dan variabel dependen (varibael yang dipengaruhi adanya varibael independen) dari data yang akan di analisis.

3. *Split Data*

Split data (membagi data) untuk membagi data menjadi data *train* dan data *set* digunakan untuk proses *training* dan *testing* dalam program. Pada penelitian ini membagi data menjadi 70% data *training* dan 30% data *testing*. Yang artinya data dari 184 data, 128 data *train* dan 56 data *test*.

4. Hitung Klasifikasi *Naïve Bayes*

Menghitung klasifikasi yang dilakukan untuk prediksi pada data train dan data test di klasifikasi *Naïve Bayes.* Setelah mendapatkan hasil prediksi dari data test selanjutnya menentukan nilai probabilitas dari data test. Setelah diperoleh nilai prediksi, maka tahap selanjutnya melakukan *Confussion Matrix*.

Confusion matrix adalah tabel yang mencatat hasil pekerjaan klasifikasi. Jumlah matriks konfusi dapat diringkas menjadi dua nilai, yaitu akurasi dan laju *error*. Dengan memahami akurasi hasil prediksi, dan dengan mengetahui jumlah data yang diklasifikasikan secara salah, kita dapat mengetahui laju *error* dari prediksi yang dibuat. Dua kuantitas ini digunakan sebagai matriks kinerja klasifikasi[12].

Perhitungan *precision* dan *recall* dapat diurutkan dengan menggunakan *confusion matrix*. Isi dari matriks konfusi adalah nilai bilangan hasil modifikasi dari keadaan nyata[9]. Anda dapat membuat dan melihat matriks pada tabel 3.2

Tabel 3.2 Komposisi Confusion Matrix

	Condition : A	Not A
Test says accepted A	True positive (TP)	False positive (FP)
Test says accepted not A	False negative (FN)	True negative (TN)

5. Hasil Klasifikasi

Dalam pengukuran performa klasifikasi ada beberapa cara, namun cara yang digunakan dalam evaluasi hasil klasifikasi program studi dengan menghitung akurasi, precision, recall dan support. Akurasi adalah persentase dari hasil semua klasifikasi yang benar dikenali. Perhitungan akurasi dilakukan dengan cara membagi jumlah data yang diprediksi benar dengan total data uji. Menghitung nilai akurasi dengan menggunakan persamaan (6)

$$Akurasi = \frac{Jumlah\ data\ yang\ diprediksi\ secara\ benar}{Jumlah\ prediksi\ yang\ dilakukan}\ x100\%$$

Persamaan 6

Metode yang digunakan selain perhitungan akurasi untuk mengukur kinerja pengklasifikasi adalah perhitungan *precission,* dan *recall*. Precision berarti hasil yang mendekati kejadian sebenarnya. Presisi ini memastikan ukuran hasil yang tepat ditentukan selama proses ekstraksi. Presisi juga merupakan metrik standar untuk menentukan apakah dokumen atau kumpulan data relevan dengan tujuan yang dimaksudkan[10].

Precission (Presisi) adalah perbandingan antara jumlah data relevan yang ditemukan dan jumlah data yang ditemukan. Presisi dihitung dengan membagi data benar dengan nilai positif dengan data benar dengan nilai positif dan data salah dengan nilai positif. Nilai data positif palsu untuk diambil dari jumlah nilai yang berbeda dari kolom positif benar yang sesuai dengan di setiap kelas[11]. Untuk menghitung nilai *precission* dapat menggunakan persamaan (7).

$$Precision = \frac{true\ positive}{true\ positive + false\ positive}$$ Persamaan 7

Perhitungan lain yang memberikan hasil yang baik adalah callback. Recall adalah ukuran pentingnya dokumen yang ada untuk hasil tertentu[10]. Recall adalah perbandingan dari jumlah materi relevan yang ditemukan terhadap jumlah materi yang relevan. Perhitugan recall dilakukan dengan cara membagi data benar bernilai positif dengan hasil penjumlahan dari data benar yang bernilai positif dan data salah yang bernilai negatif. Nilai dari data salah yang bernilai negatif diambil dari jumlah nilai selain true positive baris yang sesuai tiap kelasnya[11]. Perhitungan recall dapat menggunakan Persamaan (8).

$$Recall = \frac{true\ positive}{true\ positive + false\ negative}$$ Persamaan 8

2.7. *Logistic Regression*

Menurut Samuel Tarigan *Logistic regression* adalah bentuk khusus regresi yang diformulasikan untuk melakukan klasifikasi data ke dalam dua group (prediksi group) dan menjelaskan variabel dependen biner (kategorikal/non-metric) *Logistic regression* cocok digunakan bila kita ingin memprediksi keanggotan variabel independen (prediktor) dalam dua grup saja, misal grup orang yang menderita diabetes atau tidak menderita diabetes.

Bentuk umum logistic regression:

$$Y = X_1 + X_2 + X_3 + \ldots + X_n$$ Persamaan 9

Dimana Y merupakan *biner non-metric* dan X merupakan *non-metric* atau *metric.*

Keuntungan *logistic regression* dibandingkan dengan *multiple discriminant* analisis merupakan ketiadaan asumsi.

a. Analisis logistic regression tidak mensyaratkan
b. Bentuk distribusi tertentu buat variabel independen
c. *Homoscedasticity*
d. hubungan linier antara variabel dependen serta independen.
e. Analisis *logistic regression* mengasumsikan bahwa *independence of observations*, Linearitas dari logit (hubungan linear antara logit serta variabel independen, terutama buat yg bersifat konstan).