Skripsi Fisika

IMPLEMENTASI ALGORITMA *SUPERVISED MACHINE LEARNING* METODE NAIVE BAYES DAN K-NEAREST NEIGHBOR UNTUK MEMPREDIKSI DATA CITRA BUAH

MUHAMMAD SABRAN

H021171019

DEPARTEMEN FISIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2024

HALAMAN JUDUL

Implementasi Algoritma *Supervised Machine Learning* Metode Naive Bayes dan K-Nearest Neighbor untuk Memprediksi Data Citra Buah

SKRIPSI

Diajukan Sebagai Salah Satu Syarat Untuk Memperoleh Gelar Sarjana Sains

Pada Departemen Fisika

Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Hasanuddin

OLEH:

MUHAMMAD SABRAN

H021171019

DEPARTEMEN FISIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS HASANUDDIN

MAKASSAR

2024

HALAMAN PENGESAHAN

Implementasi Algoritma *Supervised Machine Learning* Metode Naive Bayes dan K-Nearest Neighbor untuk Memprediksi Data Citra Buah

Disusun dan Diajukan Oleh:

MUHAMMAD SABRAN

H021171019

Telah dipertahankan dan di hadapan Panitia Ujian yang dibentuk dalam rangka Penyelesaian Program Sarjana Program Studi Fisika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin

Pada 15 Mei 2024

Dinyatakan telah memenuhi syarat kelulusan

Menyetujui

Pembimbing Utama	Pembimbing Pertama
Eko Juarlin, S.Si., M.Si.	**Heryanto, S.Si., M.Si.**
NIP. 19811106 200812 1 002	**NIP. 19911129 202005 3 001**

Ketua Departemen Fisika

Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Hasanuddin

Prof. Dr. Arifin, M.T.

NIP. 19670520 199403 1 002

PERNYATAAN KEASLIAN

Saya yang bertandatangan di bawah ini:

Nama : Muhammad Sabran

NIM : H021171019

Departemen : Fisika

Judul Skripsi : Implementasi Algoritma *Supervised Machine Learning* Metode Naive Bayes dan K-Nearest Neighbor untuk Memprediksi Data Citra Buah

Menyatakan bahwa skripsi ini benar-benar hasil karya sendiri dan belum pernah diajukan untuk mendapatkan gelar sarjana di Universitas Hasanuddin atau Lembaga Penelitian lain kecuali kutipan dengan mengikuti tata penulisan karya ilmiah yang sudah lazim digunakan, karya tulis ini merupakan murni dari gagasan penelitian saya sendiri, kecuali arahan dari Tim Pembimbing dan masukan Tim Penguji.

Makassar, 15 Mei 2024

Yang membuat pernyataan

ıd Sabran

ABSTRAK

Buah-buahan menampilkan keragaman yang mencolok dalam berbagai aspek seperti warna, bentuk, luas, dan keliling, yang mencerminkan kandungan gizi dan keindahan visualnya. Klasifikasi buah menjadi integral dalam industri pertanian, pengolahan makanan, dan distribusi produk. Penggunaan pembelajaran mesin, khususnya algoritma *supervised*, semakin mendominasi dalam mengklasifikasikan buah, memberikan akurasi dan efisiensi yang tinggi. Penelitian ini bertujuan untuk menghasilkan model yang mampu mengidentifikasi data citra buah dengan menggunakan algoritma *supervised machine learning* serta menganalisis tingkat akurasinya. Metode yang digunakan adalah *K-Nearest Neighbour* (KNN) dan Naïve Bayes untuk memprediksi 128 data uji untuk 8 jenis buah dari model yang dibuat menggunakan algoritma KNN dan Naïve Bayes berdasarkan 32 data latih 8 jenis buah. Hasil penelitian menunjukkan algoritma Naive Bayes mencapai tingkat akurasi sebesar 100% dan algoritma KNN memiliki tingkat akurasi 97.66%. Sehingga model algoritma Naive Bayes menunjukkan hasil yang lebih superior dibandingkan dengan model algoritma KNN. Penelitian ini diharapkan dapat meningkatkan produktivitas, efisiensi, dan akurasi dalam berbagai aplikasi terkait buah, serta mengurangi potensi kesalahan manusia dalam proses klasifikasi.

Kata Kunci: *Supervised Machine Learning, K-Nearest Neighbour, Naïve Bayes, Model Klasifikasi, Tingkat Akurasi.*

ABSTRACT

Fruits display striking diversity in aspects such as color, shape, area, and circumference, reflecting their nutritional content and visual beauty. Fruit classification is an integral part of the agricultural industry, food processing and product distribution. The use of machine learning, especially supervised algorithms, is increasingly dominating in fruit classification, providing high accuracy and efficiency. This research aims to produce a model that is able to identify fruit image data using a supervised machine learning algorithm and analyze its level of accuracy. The method used is K-Nearest Neighbor (KNN) and Naïve Bayes to predict 128 test data for 8 types of fruit from a model created using the KNN and Naïve Bayes algorithms based on 32 training data for 8 types of fruit. The research results show that the Naive Bayes algorithm achieved an accuracy level of 100% and the KNN algorithm had an accuracy level of 97.66%. So the Naive Bayes algorithm model shows superior results compared to the KNN algorithm model. This research is expected to increase productivity, efficiency and accuracy in various fruit-related applications, as well as reduce human potential in the classification process.

Keywords: *Supervised Machine Learning, K-Nearest Neighbor, Naïve Bayes, Classification Model, Level of Accuracy.*

KATA PENGANTAR

Assalamu'alaikum Wa Rahmatullahi Wa Barakaatuuh.

Puja dan puji syukur penulis panjatakan kepada Allah SWT berkat limpahan rahmat nikmat dan karunia-Nya sehingga penulis dapat menyelesaikan penyusunan skripsi dengan dengan judul "Implementasi Algoritma *Supervised Machine Learning* Metode Naive Bayes dan K-Nearest Neighbor untuk Memprediksi Data Citra Buah" yang merupakan salah satu syarat untuk mendapatkan gelar sarjana sains di Departemen Fisika Program Studi Fisika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin. Serta salawat dan salam terus tercurahkan oleh penulis terhadap manusia yang menjadi sosok serta suri tauladan bagi umat manusia di muka bumi ini baginda Rasulullah Muhammad SAW.

Selama proses penyelesaian skripsi, penulis mengalami berbagai hambatan dan menyadari bahwa skripsi ini masih jauh dari kesempurnaan. Hambatan dapat teratasi tentu tidak lepas dari bimbingan, dukungan, dan bantuan dari berbagai pihak. Oleh karena itu, penulis mengucapkan terimakasih yang sebesar-besarnya kepada orang-orang yang turut membantu, baik dalam bentuk sumbangan ide, materil, maupun moril sehingga skripsi ini dapat selesai sebagaimana mestinya. Dengan segala kerendahan hati penulis mengucapkan banyak terima kasih yang sebesar-besarnya kepada:

1. Keluarga tercinta, terkhusus kepada kedua orang tua tercinta, Yth. Ayah **Basri Hapil** dan Ibu **Rahmaniah** yang selalu memberikan kasih sayang, perhatian, semangat, dan dukungan baik secara moral maupun secara materi kepada penulis, dan ucapan terima kasih penulis sampaikan kepada adik-adik **Muh. Radian**, **Ardina Rahma**, dan **Aprilia Aisyah** yang selalu memberikan dukungan dalam bentuk apapun itu.
2. Tante sekaligus wali bagi penulis, Yth. Tante **Maryam, S.Pd** dan Om **Akhiruddin, S.Pd** yang menjadi sosok berpangaruh dalam hidup saya, selalu mengingatkan, memberikan dukungan baik secara moral maupun secara materi kepada penulis.

3. Bapak **Prof. Dr. Arifin, M.T.** selaku ketua Departemen Fisika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin dan juga selaku dosen pembimbing akademik yang telah membantu dan memberikan banyak saran dan nasehat yang membuat penulis merasa tertolong selama menjalani masa perkuliahan.
4. Bapak **Eko Juarlin, S.Si., M.Si.** selaku Pembimbing Utama dan Bapak **Heryanto, S.Si., M.Si.** selaku Dosen Pembimbing Pertama yang telah banyak membimbing dan meluangkan waktu, tenaga, serta pemikirannya untuk penulis sehingga skripsi ini dapat terselesaikan.
5. Bapak **Prof. Dr. Tasrief Surungan, M.Sc.** dan **Dr. Sri Dewi Astuty, S.Si., M.Si.** selaku Tim Penguji yang telah banyak meluangkan waktu dan tenaga untuk memberikan ilmu, saran, dan diskusi dalam menyelesaikan skripsi ini.
6. Seluruh **Dosen FMIPA Unhas**, khususnya kepada seluruh Bapak dan Ibu **Dosen Pengajar Departemen Fisika**, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Hasanuddin. Terima kasih telah memberikan ilmu yang bermanfaat dan mendidik penulis selama menjadi mahasiswa di kampus merah ini
7. Bapak/Ibu Staf Pegawai FMIPA UNHAS, terutama kepada Pak **Suardi** serta Staf Departemen Fisika, terutama Bu **Rana,** Bu **Evi,** dan Pak **Syukur** yang selalu membantu penulis dalam mengurus berkas selama berada di kampus.
8. Kawan-kawan pengurus BEM FMIPA Unhas Periode 2020/2021 yang telah banyak menguras tenaga dan pikiran selama menjalankan kepengurusan. Kepada ketua BEM **Rahman** terimakasih sudah dipercayakan sebagai anggotanya.
9. 79 orang yang awalnya tak kukenal kemudian menjadi Saudara meskipun tak sedarah, **Himafi17 (Faqih, Aat, Khalis, Callu, Fajar, Madan, Agung, Qoil, Jepri, Wahyu, Ardi, Zain, Ucha, Ebiet, Dandung, Ale, Roni, Gabe, Fadlan, Faishal, Tsaqif, Ano, Reza, Indra, Adi, Aldo, Angga, Puad, Bintang, Zahari, Rial, Albaar, Nurul Fauziah, Nia, Esi, Ainun, Syakirah, Epi, Khusnul, Riri, Kiki, Melsi, Daya, Miftah, Mayama, Wide, Desha, Eky, Illa, Gebrina, Yusrin, Evita, Egi, Sappe, Mirna, Hikmah, Destri, Ate, Asni,**

Suci, Ghufa, Titien, Nova, Rachel, Ningnang, Gita, Lahu, Yesi, Rapang, Adhe, Unia, Danti, Cammai, Time, Ola, Rahmah, Ajeng, Sindy, Cindy) terima kasih atas kepercayaannya kepada saya untuk menjadi Ketua di Angkatan Kita, Terimakasih juga untuk kesadaran, kebersamaan, keselarasan, kepusingan dan segala bentuk dukungan yang kalian berikan. Kalian orang hebat yang selalu ada baik suka maupun duka. Saya sangat bersyukur bisa berada dan memimpin kalian. Semoga tetap Teguh dalam Keyakinan, Kukuh dalam Kebersamaan

10. Para Tampan Maks, **Zahari, Puad, Faqih, Angga, Aat, Callu, Gabe, Fajar, Madan, Agung, Qoil, Jepri, Wahyu, Ardi, Zain, Ucha, Ebiet, Dandung, Ale, Roni, Fadlan, Faishal, Tsaqif, Ano, Reza, Indra, Adi, Aldo, Khalis, Rial, Albaar**. Terima kasih telah menjadi anomali yang nyata di mipa (kehadiran cowok lebih banyak daripada cewek).
11. Kawan-kawan **MIPA 2017**, mulai dari Mipa 1 sampai Mipa 7 secara keseluruhan. Terutama Ketua Angkatan (yang menang suara dari saya), **Roni Rahmat** terimakasih telah merangkul dan bermusyawarah kepada semua anggota MIPA 2017 dan tetap memahamkan slogan KAMI SATU KAMI BERSAUDARA.
12. Kanda – kanda Warga Himafi terutama Kanda – Kanda Himafi 2015 selaku Pengurus Himpunann kami dan Kanda – kanda Himafi 2016 selaku panitia kami yang telah banyak memberikan arahan dan masukan dan arahan selama saya menjadi mahasiswa, baik akademik maupun non-akademik.
13. Adik – adik tak sedarah, Himafi 2018 (Dede, Fauzi, justin, tara, gopal, azlan, sarwan, heral, ipul, hasnan, boca, micin, jihan, via, fira, zefa, sorong, dena dan adik-adik yang belum sempat disebutkan namanya).
14. Adik-adik Himafi dan HMGF 2019 (Yusri, Stefen, Agung, Alya, Nurul, Kopat, Akbar Nabawi Faturrahman, Haikal, Alif, Mawang, Haidir, Arsyi, Haerul, Reika, Cindy, Ita, Devi, Sarni, Fatihah, Mey, Maulidah, Nude, Nismul, Ismi, Dian dan adik-adik lain yang belum sempat disebutkan Namanya).

15. MIPA 2020 (Algi, Toktok, Asmawan, Danke, Site, Vikram, Letnan, Iskar, dan yang lainnya) terima kasih atas dukungan, semangat yang pernah ada dan semoga selalu ada. Semoga tetap Satu dan Selamanya
16. Lembagaku, **Himpunan Mahasiswa Fisika (Himafi)** FMIPA Unhas terima kasih telah membentuk karakter keras, kuat, cerdas dan berani di dalam diri penulis, serta memperkenalkan dan mengajarkan banyak hal baru sejak penulis menjadi mahasiswa baru hingga saat ini.
17. Adik-adik 2021 (Palele, Suliz, Reynol, Sute, Karappe, Sarah, Rifkah, Bobo, Mar'ah, Fera, Patra, Vadya, Amar, Ulfa, Kiki, Ulia, Nanda, Ici, dan adik-adik yang belum sempat disebutkan namanya).
18. Kawan-kawan KKN Takalar 11 Kabs Only (Arya, Alim, Viqri, Rezky).
19. Teruntuk Muh. Zahary, terima kasih telah menjadi teman sekaligus saudara yang kurang tidur namun mempunyai banyak mimpi.
20. Teruntuk Puad Ary Prasetya, terima kasih telah menjadi teman sekaligus saudara yang selalu memberi support dan bantuan dalam menyelesaikan skripsi ini.
21. Semua pihak yang tidak dapat penulis sebutkan satu persatu, yang telah memberikan kontribusi sehingga skripsi ini dapat terselesaikan dengan baik.

Makassar,12 Oktober 2022

Muhammad Sabran

DAFTAR ISI

HALAMAN JUDUL ii
HALAMAN PENGESAHAN iii
PERNYATAAN KEASLIAN iv
ABSTRAK v
ABSTRACT vi
KATA PENGANTAR viii
DAFTAR ISI xii
DAFTAR GAMBAR 1
DAFTAR TABEL 1
BAB I PENDAHULUAN 2
I.1 Latar Belakang 2
I.2 Rumusan Masalah 3
I.3 Tujuan Penelitian 3
BAB II TINJAUAN PUSTAKA 4
II.1 Pembelajaran Mesin (*Machine Learning*) 4
II.1.1 Naive Bayes 4
II.1.2 K-Nearest Neighbour 7
II. 2 RapidMiner 9
II. 3 Pengolahan Citra Digital 10
II.4. Kaggle 11
BAB III METODOLOGI PENELITIAN 12
III.1 Prosedur Penelitian 12
III.2 Pengolahan Citra Digital dengan Octave 12
III.3 Naive Bayes (See Rapidminer) 13
III.4 KNN (See Rapidminer) 14
III.5 Analisis 14
III.6 Bagan Alir Penelitian 15
BAB IV HASIL DAN PEMBAHASAN 16
IV.1 Analisis *Machine Learning* RapidMiner 16
IV.1.1 *Exploratory Data Analysis* (EDA) 16
IV.1.2 Data Mining Process 17

IV.1.2 Evaluasi dengan Matriks Konfusi 18
IV.2 Pembahasan Hasil Penelitian 20
BAB V PENUTUP 21
V.1 Kesimpulan 21
V.2 Saran 21
DAFTAR PUSTAKA 22
LAMPIRAN 25
Lampiran 1. Gambar Perangkat Lunak Octave dan RapidMiner 25
Lampiran 2. Gambar *Workspace* pada Octave Setelah Program Dijalankan 26
Lampiran 3. Gambar Dataset pada Situs Kaggle 27
Lampiran 4. Data Citra Sebelum Normalisasi 28
Lampiran 5. Data Citra Setelah Normalisasi 34

DAFTAR GAMBAR

Gambar 2.1. 8 jenis buah dari dataset kaggle; a. Apel; b. Pisang; c. Kiwi;.....11
Gambar 4.1. Tampilah Exploratatory Data Analysis (EDA). 16
Gambar 4.2. Tahap awal proses analisis data latih dan data uji (kiri). Parameter pada widget split data (kanan)..... 17
Gambar 4.3. Tampilan widget split data saat memasukkan rasio. 17
Gambar 4.4. Tahapan proses analisis data menggunakan model algoritma. 18
Gambar 4.5. Nilai matriks konfusi model Naive Bayes. 19
Gambar 4.6. Nilai matriks konfusi KNN. 19

DAFTAR TABEL

Tabel 2.1. Contoh dataset play tennis (UCI machine learning repository)..... 6
Tabel 2.2. Frekuensi setiap nilai attribut..... 6
Tabel 2.3. Probabilitas setiap nilai attribut. 6
Tabel 2.4. Contoh testing data play tennis. 6
Tabel 2.5. Dataset wilayah dalam koordinat kartesian 8
Tabel 2.6. Data tanpa daerah untuk data uji. 8
Tabel 2.7. Tabel jarak data uji terhadap data latih. 8

BAB I
PENDAHULUAN

I.1 Latar Belakang

Buah-buahan memiliki keragaman yang mencolok, dibedakan oleh sejumlah faktor seperti warna, bentuk, luas, dan keliling. Salah satu aspek perbedaan yang menonjol adalah warna buah, yang tidak hanya memberikan keindahan visual, tetapi juga mencerminkan kandungan vitamin yang dimilikinya. Dengan demikian, melalui observasi terhadap atribut-atribut ini, kita dapat memahami lebih dalam tentang keanekaragaman alam dan manfaat gizi yang dapat diperoleh dari berbagai jenis buah. Perbedaan ini memiliki keterkaitan yang erat dengan bidang taksonomi, di mana karakteristik unik setiap buah membantu dalam klasifikasi dan identifikasi jenisnya.

Untuk mengklasifikasikan buah, penggunaan pembelajaran mesin menjadi semakin mendominasi karena memberikan keunggulan dalam akurasi dan efisiensi. Klasifikasi buah adalah bagian integral dari berbagai industri, termasuk pertanian, pengolahan makanan, dan distribusi produk. Pembelajaran mesin menawarkan pendekatan yang terstruktur dan terukur untuk memahami dan memanfaatkan pola-pola kompleks yang mungkin sulit diidentifikasi secara manual.

Pembelajaran mesin memiliki dua tipe algoritma: algoritma *supervised* dan algoritma *unsupervised*. Salah satu keunggulan utama pembelajaran mesin algoritma *supervised* dalam konteks ini adalah kemampuannya untuk mencapai tingkat akurasi yang tinggi. Studi oleh Zhang dkk[1]. (2022) dalam "*Computers and Electronics in Agriculture*" menggambarkan penerapan pembelajaran mesin algoritma *supervised* untuk memamanen dan menyortir jamur secara otomatis berdasarkan citra digital.

Pembelajaran mesin algoritma *supervised* memanfaatkan data berlabel selama fase pelatihan, memungkinkan model untuk memahami pola-pola yang terkait dengan berbagai jenis buah dan jenis benda lainnya. Penelitian oleh Gupta dkk[2]. (2023) dalam "*Computational Materials Science*" mengadopsi pendekatan pembelajaran mesin algoritma *supervised* untuk mengklasifikasikan struktur baja

mikro berdasarkan karakteristik data struktur baja mikro, menunjukkan bahwa penggunaan data berlabel menghasilkan model yang lebih akurat dan adaptif.

Pada algoritma pembelajaran mesin terdapat enam teknik pembelajaran mesin terbimbing yang digunakan dalam *multiclass classification* termasuk K *Nearest Neighbour* (KNN), *Support VectorMachine*, Naïve Bayes, Analisis Diskriminan Linier, Pohon Keputusan, dan jaringan saraf propagasi maju umpan balik.

Pembelajaran mesin algoritma *supervised* juga memungkinkan inovasi dalam mengatasi tantangan klasifikasi multikelas, di mana terdapat banyak jenis buah yang perlu diidentifikasi. Penelitian oleh Gupta dkk[2]. (2023) dalam "*Computational Material Science*" mengeksplorasi penggunaan algoritma *K-Nearest Neighbour* untuk mengklasifikasikan tipe struktur mikro baja dalam skenario multikelas, menghasilkan pemisahan kategori yang jelas dan efisien.

Berdasarkan hal yang telah dijelaskan di atas, dilakukan penelitian ini guna untuk mengklasifikasikan jenis buah menggunakan pembelajaran mesin algoritma *supervised* metode Naive Bayes dan *K-Nearest Neighbour*. Pada penelitian ini diharapkan mampu membantu meningkatkan produktivitas, efisiensi, dan akurasi dalam berbagai aplikasi terkait buah, serta meminimalkan potensi kesalahan manusia dalam proses klasifikasi.

I.2 Rumusan Masalah

Rumusan masalah pada penelitian ini adalah:

1. Bagaimana mengidentifikasi data citra delapan jenis buah dari dataset Kaggle dengan metode KNN dan Naive Bayes.
2. Berapa tingkat akurasi dari penerapan metode KNN dan Naive Bayes.

I.3 Tujuan Penelitian

Tujuan penelitian ini adalah:

1. Menghasilkan model yang mampu mengidentifikasi data citra buah menggunakan metode KNN dan Naive Bayes.
2. Menganalisis akurasi dari penerapan metode KNN dan Naive Bayes.

BAB II
TINJAUAN PUSTAKA

II.1 Pembelajaran Mesin (*Machine Learning*)

Pembelajaran mesin adalah metode kecerdasan buatan yang penting yang digunakan dalam penelitian. Metode ini berupaya mempelajari pengetahuan dan aturan dari data yang kompleks untuk memprediksi hasil dan perilaku di masa depan. Berbeda dengan metode statistik, pembelajaran mesin lebih berfokus untuk mendapatkan pencapaian prediksi aktual serta kinerja prediksi yang lebih baik[3]. Berdasarkan metode pembelajaran yang digunkan, pembelajaran mesin dapat dibagi menjadi pembelajaran mesin terbimbing (*supervised*) dan pembelajaran mesin tanpa pengawasan (*unsupervised*). Pembelajaran mesin terbimbing (*Supervised Machine Learning*) umumnya digunakan untuk menyelesaikan tugas regresi dan klasifikasi. Input dan output pada metode ini ditentukan di awal kemudian mesin mempelajari pola dari input hingga output yang diharapkan[4] Metode klasifikasi yang sering digunakan dalam pembelajaran mesin yaitu:

II.1.1 Naive Bayes

Ilmuwan asal Inggris Thomas Bayes awalnya mengemukakan Teorema Bayes yang menyatakan bahwa prediksi kemungkinan-kemungkinan di masa depan dapat ditentukan berdasarkan pengalaman masa lalu yang sekarang dikenal sebagai Naive Bayes[5]. Naive Bayes adalah salah satu teknik klasifikasi algoritma supervised machine learning yang menggunakan metode probabilistik dan statistik untuk membagi data ke dalam kelas (himpunan) berdasarkan ciri masing-masing. Naive Bayes *classifier* adalah salah satu teknik klasifikasi sederhana dan efektif yang membantu menciptakan model pembelajaran mesin cepat yang dapat membuat prediksi dengan cepat[6].

Probabilitas korespondensi fitur-fiturnya terhadap kelas masing-masing ditentukan dengan melakukan beberapa tahapan, yaitu menghitung probabilitas setiap kelas dengan masing-masing fiturnya, melakukan normalisasi dari hasil yang didapatkan, kemudian memilih kelas dengan probabilitas tertinggi.

$$P(C_i|f_1, f_2, \dots, f_n) \propto \frac{P(C_i) \times \prod_{j=1}^{n} P(f_j|C_i)}{P(f_1) \times P(f_2) \dots \times P(f_n)} \quad (2.1)$$

Atau

$$\text{likelihood}(c_i) = P(c_i) \prod_{f=1}^{F} P(t_f|c_i) \quad (2.2)$$

C_i = nilai kelas ke i dari no. kelas C

f_j = fitur ke-j dari kumpulan data.

Probabilitas dari setiap kelas dapat dihitung setelah menghitung frekuensi nilai atribut dari masing-masing kelas. Di sini, Persamaan (2.2) menunjukkan probabilitas setiap kelas dengan masing-masing fitur yang diberikan, yang berbanding lurus dengan produk probabilitas setiap fitur; ketika kelasnya Ci terhadap probabilitas suatu kelas, dan ini dibagi dengan hasil kali probabilitas setiap fitur[7].

Untuk menghindari redudansi data dan munculnya fitur yang tidak diinginkan maka perlu dilakukan normalisasi data (Persamaan (2.3)). Menetukan output dari record tertentu dilakukan dengan menambahkan argmax, yang akan memberi kita probabilitas tertinggi dari kelas tersebut seperti yang ditunjukkan pada Persamaan. (2.4), yang paling mungkin[6].

$$P_{assignment}(c_i) = \frac{\text{likelihood}(c_i)}{\sum_{c_j \epsilon C} \text{likelihood}(c_j)} \quad (2.3)$$

$$c = argmax_{i=1,2,..k} P(C_i) \times \prod_{j=1}^{n} P(f_j|C_i) \quad (2.4)$$

Untuk mengklasifikasikan data berdasarkan algoritma Naive Bayes dilakukan dengan beberapa tahapan. Pertama, menghitung probabilitas setiap kelas berdasarkan frekuensi dari masing-masing fitur yang diberikan menggunakan persamaan 2.1. Misalnya pada Tabel 2.2 diberikan frekuensi setiap nilai attribut dari dataset bermain tenis berdasarkan Tabel 2.1. Model probabilitas berdasarkan frekuensi diberikan pada Tabel 2.3[8].

Tabel 2.1. Contoh dataset play tennis (UCI machine learning repository).

id	outlook	temperature	humidity	windy	play (class)
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

Tabel 2.2. Frekuensi setiap nilai attribut.

	outlook			temperature			humidity			windy		play (class)	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	3	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								

Tabel 2.3. Probabilitas setiap nilai attribut.

	outlook			temperature			humidity			windy		play (class)	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2/9	3/5	hot	2/9	3/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

Untuk menguji kebenaran dari model yang telah dibangun berdasarkan algoritma Naive Bayes, digunakan testing data (Tabel 2.4).

Tabel 2.4. Contoh testing data play tennis.

id	outlook	temperature	humidity	windy	play (class)
1	sunny	cool	high	true	no

$$\begin{aligned}\text{likelihood}(play = yes) &= P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes) \\ &\quad P(true \mid yes) \\ &= \frac{9}{14} * \frac{2}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} \\ &= 0.0053\end{aligned}$$

$$\begin{aligned}\text{likelihood}(play = no) &= P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no) \\ &\quad P(true \mid no) \\ &= \frac{5}{14} * \frac{3}{5} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5} \\ &= 0.0206\end{aligned}$$

$$\begin{aligned}P_{assignment}(play = yes) &= \frac{\text{likelihood}(play = yes)}{\text{likelihood}(play = yes) + \text{likelihood}(play = no)} \\ &= \frac{0.0053}{0.0053 + 0.0206} \\ &= 0.205\end{aligned}$$

$$\begin{aligned}P_{assignment}(play = no) &= \frac{\text{likelihood}(play = no)}{\text{likelihood}(play = yes) + \text{likelihood}(play = no)} \\ &= \frac{0.0206}{0.0053 + 0.0206} \\ &= 0.795\end{aligned}$$

Karena $P_{assignment}(play = no) > P_{assignment}(play = yes)$ maka diputuskan bahwa kelas untuk data uji tersebut adalah $play = no$.

II.1.2 K-Nearest Neighbour

Teknik pembelajaran mesin terbimbing yang disebut *K-Nearest Neighbour* (KNN) digunakan untuk mengklasifikasikan berdasarkan jangkauan tetangga terdekatnya ketika cakupan tetangga terdekat diatur ke K[9]. Nilai default K diambil berdasarkan jumlah keseluruhan dari kelas data yang digunakan ditambah satu. Matriks jarak yang digunakan untuk model adalah jarak garis lurus yang umum digunakan (Persamaan (2.5)) untuk data kontinu[10].

$$d_{euclidean} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \tag{2.5}$$

Data jarak dari tetangga yang didapatkan berdasarkan Persamaan (2.5) kemudian diurutkan berdasarkan jarak yang terkecil beserta kelas yang berkaitan. K tetangga terdekat dari hasil tersebut dipilih sesuai dengan nilai K yang telah ditentukan. Kelas terbanyak dari K tetangga terdekat itu menjadi kelas bagi sampel tersebut.

Untuk mengklasifikasikan data berdasarkan algoritma Naive Bayes dilakukan dengan menghitung jarak tetangga terdekat dari data uji berdasarkan Persamaan (2.5). Misalnya, menentukan letak attribut wilayah dari data uji pada Tabel 2.6 berdasarkan dataset wilayah untuk atribut pada koordinat x dan y seperti pada Tabel 2.5.

Tabel 2.5. Dataset wilayah dalam koordinat kartesian

X	Y	Daerah
1	2	KOTA
2	1	KOTA
3	-1	KOTA
4	0	KOTA
3	2	KOTA
2	-2	KOTA
5	0	KAB
1	-4	KAB
2	5	KAB
6	1	KAB
6	-1	KAB
7	-3	KAB
2	6	KAB

Tabel 2.6. Data tanpa daerah untuk data uji.

X	Y	Daerah
3	3	?

Setelah mendapatkan data jarak dari masing-masing kelas terhadap data uji, dilakukan pengurutan data dari nilai yang terkecil hingga terbesar beserta atribut daerahnya (Tabel 2.7). Didapatkan nilai attribut daerah berdasarkan nilai K yang telah ditentukan yaitu Kota.

Tabel 2.7. Tabel jarak data uji terhadap data latih.

Daerah	Dis
KOTA	1,41
KOTA	2,00
KAB	2,24
KOTA	3,16
KAB	3,61
KAB	3,61
KAB	4,00
KOTA	4,12

KOTA	5,00
KAB	5,00
KOTA	5,39
KAB	8,06
KAB	9,06

II. 2 RapidMiner

RapidMiner adalah alat analitik berbasis Java dari ujung ke ujung yang digunakan untuk penambangan data, penambangan teks, analisis prediktif, dan analisis bisnis yang dikembangkan oleh perusahaan dengan nama yang sama. RapidMiner merupakan solusi yang telah digunakan di beberapa bidang dan sebenarnya merupakan solusi mandiri dan sumber terbuka yang paling populer, serta pemimpin pasar di bidangnya[11]. Alat ini dapat mengakses lebih dari 40 jenis file termasuk SAS, ARFF, Stata dan melalui URL; akses ke dokumen teks, halaman web, PDF, HTML dan XML, serta database NoSQL, MongoDB dan Cassandra; model dengan serangkaian kemampuan pemodelan dan algoritme yang lebih besar seperti penghitungan kesamaan, pengelompokan, analisis keranjang pasar, pohon keputusan, induksi aturan, pemodelan Bayesian, regresi, jaringan saraf, mesin vektor dukungan, penalaran berbasis memori, ansambel model, dan estimasi kinerja model dengan beberapa teknik validasi dan kriteria kinerja; dan menganalisis data dalam jumlah besar melalui beberapa alat lain seperti Hadoop, Spark, Hive, MapReduce, Pig, dan Mahout[12].

Solusi dari RapidMiner memiliki lima alat/edisi: i) RapidMiner Studio, aplikasi klien yang mendukung penerapan alur kerja analitik prediktif lengkap dengan fitur VCF yang dapat mencakup tugas utama penambangan data seperti data integrasi, pembersihan, transformasi, eksplorasi, pemodelan, dan validasi dengan *graphical user interface* (GUI); ii) RapidMiner Server, server untuk menjalankan pekerjaan otomatis dan terjadwal, kerja tim kolaboratif, pembuatan aplikasi berbasis web, serta penerapan dan integrasi dengan sistem lain; iii) RapidMiner Radoop, untuk mempercepat analisis data dalam jumlah besar dan mengatasi kompleksitas yang dimiliki Spark Hadoop pada pengguna non-teknis dengan menggunakan serangkaian kemampuan untuk melakukan penambangan data di Hadoop; iv) Ekstensi RapidMiner, kemampuan tambahan yang disediakan oleh

komunitas seperti Pemrosesan Teks, Penambangan Web, Ekstensi WeKa, Analisis Teks oleh AYLIEN, dan Ekstensi Seri; dan v) RapidMiner Cloud, kemampuan yang memungkinkan pekerjaan penambangan data diproses di cloud. RapidMiner Studio didukung oleh Windows, Mac OS dan Linux[12].

II. 3 Pengolahan Citra Digital

Pengolahan citra adalah pemrosesan berdasarkan citra, khususnya dengan menggunakan komputer, menjadikan citra dengan kualitas yang lebih baik. Pengolahan citra bertujuan untuk memperbaiki kualitas citra agar mudah diintreprestasikan oleh manusia atau mesin secara cepat dan akurat. Dengan menganalisis citra visual buah, sistem dapat membedakan antara jenis buah yang berbeda, memilah buah yang cacat, dan mengklasifikasikan berbagai varietas berdasarkan fitur visual tertentu[13].

Proses pengolahan citra dimulai dari perubahan warna menjadi angka warna RGB (*red, green blue*) atau *grayscale*. Dari angka warna, dihitung keliling objek. Penghitungan keliling bisa dilakukan dengan operator Sobel. Dari angka warna luas dapat dihitung dengan teknik pengambangan. Pemilihan parameter ini tergantung pada tujuan akuisisi dan karakteristik objek yang diambil gambar.

Keliling objek bisa dihitung dengan menggunakan operator Sobel dan pengambangan. Operator Sobel merupakan operator klasik untuk deteksi tepi orde pertama yang melakukan pengukuran gradien spasial 2D pada gambar dan umumnya digunakan untuk mencari perkiraan besaran gradien absolut pada setiap piksel[14]. Perhitungan dengan menggunakan operator Sobel deteksi tepi untuk mengetahui keliling pada gambar. Operator Sobel memiliki dua kernel untuk konvolusi vertikal dan horizontal pada gambar hipotetis. Inti-inti ini, dilambangkan dengan Gx dan Gy, direpresentasikan pada Persamaan (2.6) dan (2.7). Konvolusi terjadi antara gambar dan kernel yang bersangkutan untuk menampilkan piksel tepi[15].

$$G_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} \tag{2.6}$$

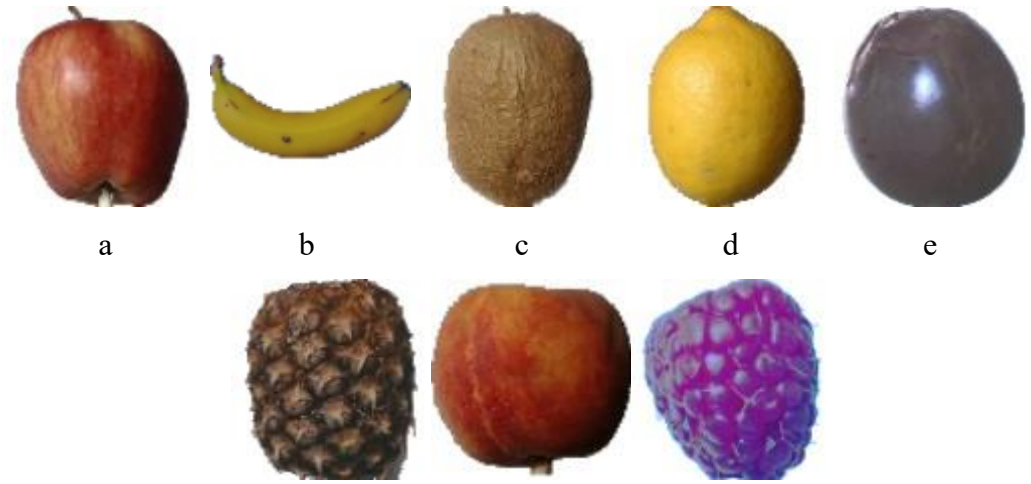
$$G_x = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} \tag{2.7}$$

Luas dengan metode pengambangan dalam proses pengolahan citra digital digunakan untuk menghitung luas dari citra objek 2D berdasarkan hasil konversi RGB ke ruang warna *gryscale*[16].

II.4. Kaggle

Kaggle adalah platform komunikasi online di mana peserta dapat menghasilkan konten dan berinteraksi dengan pengguna lain. Peserta dapat membentuk komunitas virtual seperti buku catatan online dan forum untuk mendiskusikan pilihan model, fitur, fungsi kerugian, dan sebagainya. Ketersediaan konten buatan pengguna yang kaya dari platform Kaggle memungkinkan untuk mengkaji fenomena ini dari perspektif analitis kuantitatif menggunakan penambangan teks dan analisis media sosial. Dikombinasikan dengan analisis kualitatif seperti studi kasus, dapat memberikan wawasan dan meningkatkan pemahaman yang diperoleh melalui analisis media sosial[17].

Dataset dari kaggle yang dipublikasikan oleh Tri Do memuat citra dari 8 jenis buah yang berjudul "8fruits" dapat dilihat pada Gamber 2.1[18]. Data tersebut kemudian diolah menggunakan beberapa metode pengolahan citra digital dengan bantuan perangkat lunak dari komputer.



Gambar 2.1. 8 jenis buah dari dataset kaggle; a. Apel; b. Pisang; c. Kiwi; d. Lemon; e. Manggis; f. Nanas; g. Persik; h. Frambos.