

**PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* PADA DATA
KATEGORIK DENGAN PERHITUNGAN JARAK MENGGUNAKAN
*WEIGHTED SIMPLE MATCHING COEFFICIENT***

(Skripsi)

Oleh

**CITRA PUSPA TRIA
NPM. 1917031074**



**JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023**

ABSTRACT

APPLICATION OF K-NEAREST NEIGHBOR ALGORITHM TO CATEGORICAL DATA BY CALCULATING DISTANCE USING WEIGHTED SIMPLE MATCHING COEFFICIENT

By

CITRA PUSPA TRIA

K-Nearest Neighbor is one of classification algorithm that classifies objects based on the majority class of k-nearest objects. Commonly, the measure of proximity between objects is calculated using Euclidean distances. However, if the data being used is categorical, Euclidean distances are seen as improper to apply. The weighted simple matching coefficient (WSMC) method, which calculates how close two objects are to one another, is one solution to this issue. This research was conducted to build a classification model for categorizing laptop pricing ranges based on 4 features offered, namely laptop brand, processor, RAM capacity, and storage capacity. The classification model is built on two proportion of training data and testing data, which is 80:20 and 90:10. Based on the analysis's results, the model with a 90:10 data ratio was found to be the best classification model, with accuracy values of 86.96%, recall of 50%, precision of 66.67%, and f1-score of 57.14%.

Keywords: Classification, K-Nearest Neighbor (KNN), Weighted Simple Matching Coefficient (WSMC)

ABSTRAK

PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* PADA DATA KATEGORIK DENGAN PERHITUNGAN JARAK MENGGUNAKAN *WEIGHTED SIMPLE MATCHING COEFFICIENT*

Oleh

CITRA PUSPA TRIA

K-Nearest Neighbor merupakan salah satu algoritma klasifikasi yang mengklasifikasikan objek berdasarkan kelas mayoritas dari k objek terdekat di sekitarnya. Pada umumnya, ukuran kedekatan antar objek dihitung dengan menggunakan jarak Euclidean. Akan tetapi, jarak Euclidean dirasa kurang tepat digunakan apabila data yang digunakan adalah data kategorik. Salah satu alternatif dalam mengatasi permasalahan tersebut adalah dengan menggunakan metode *weighted simple matching coefficient* (WSMC) sebagai ukuran kedekatan antar objek. Penelitian ini dilakukan untuk membangun model klasifikasi untuk penentuan kelas rentang harga laptop berdasarkan 4 fitur yang ditawarkan, yaitu merek laptop, prosesor, kapasitas RAM, dan kapasitas penyimpanan. Model klasifikasi dibangun berdasarkan dua proporsi pembagian data latih dan data uji, yaitu 80:20 dan 90:10. Berdasarkan hasil analisis, diperoleh model klasifikasi terbaik, yaitu model dengan proporsi 90% data latih dan 10% data uji, dengan nilai akurasi sebesar 86.96%, *recall* sebesar 50%, presisi sebesar 66.67%, dan *f1-score* sebesar 57.14%.

Kata Kunci: Klasifikasi, *K-Nearest Neighbor* (KNN), *Weighted Simple Matching Coefficient* (WSMC)

**PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* PADA DATA
KATEGORIK DENGAN PERHITUNGAN JARAK MENGGUNAKAN
WEIGHTED SIMPLE MATCHING COEFFICIENT**

Oleh

CITRA PUSPA TRIA

Skripsi

**Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA MATEMATIKA**

Pada

**Jurusan Matematika
Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung**



**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023**

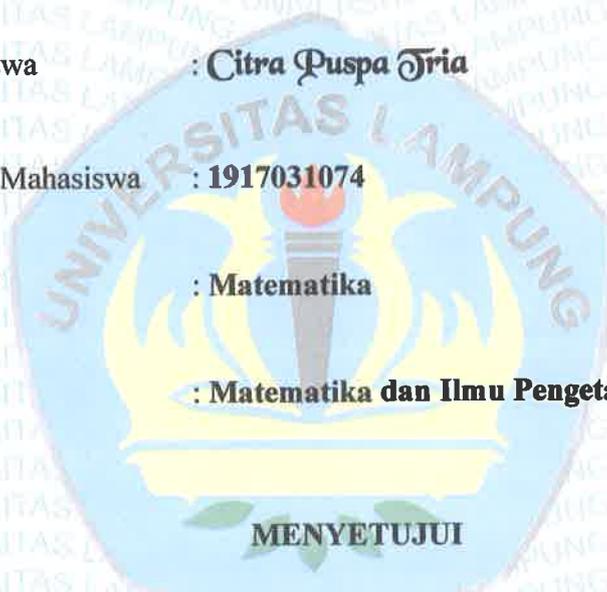
Judul Skripsi : **PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* PADA DATA KATEGORIK DENGAN PERHITUNGAN JARAK MENGGUNAKAN *WEIGHTED SIMPLE MATCHING COEFFICIENT***

Nama Mahasiswa : **Citra Puspa Tria**

Nomor Pokok Mahasiswa : **1917031074**

Program Studi : **Matematika**

Fakultas : **Matematika dan Ilmu Pengetahuan Alam**



1. Komisi Pembimbing

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 197403162005011001

Dr. Ahmad Faisol, S.Si., M.Sc.
NIP. 198002062003121003

2. Ketua Jurusan Matematika

Dr. Aang Nuryaman, S.Si., M.Si.
NIP. 197403162005011001

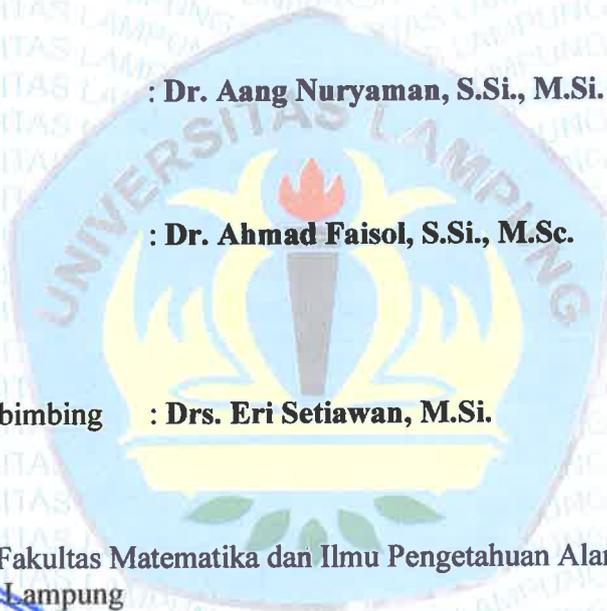
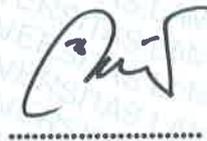
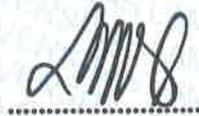
MENGESAHKAN

1. Tim Penguji

Ketua : Dr. Aang Nuryaman, S.Si., M.Si.

Sekretaris : Dr. Ahmad Faisol, S.Si., M.Sc.

**Penguji
Bukan Pembimbing : Drs. Eri Setiawan, M.Si.**



**2. Plt. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung**



Dr. Eng. Heri Satria, S.Si., M.Si.
NIP. 197110012005011002

Tanggal Lulus Ujian Skripsi: 28 Maret 2023

PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama : **Citra Puspa Tria**
Nomor Pokok Mahasiswa : **1917031074**
Jurusan : **Matematika**
Judul Skripsi : **Penerapan Algoritma *K-Nearest Neighbor* pada Data Kategorik dengan Perhitungan Jarak Menggunakan *Weighted Simple Matching Coefficient***

Dengan ini menyatakan bahwa penelitian ini adalah hasil pekerjaan saya sendiri dan apabila di kemudian hari terbukti bahwa skripsi ini merupakan hasil Salinan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan ketentuan akademik yang berlaku.

Bandar Lampung, 11 April 2023
Yang Menyatakan,



Citra Puspa Tria
NPM. 1917031074

RIWAYAT HIDUP

Penulis bernama Citra Puspa Tria yang lahir di Kota Bandar Lampung pada tanggal 29 April 2002. Penulis lahir dari pasangan Dedi dan Aike Adriani dan merupakan anak ketiga dari empat bersaudara.

Penulis memulai pendidikan pada tahun 2007 di Taman Kanak-Kanak Ar-Rusydah III. Kemudian, melanjutkan pendidikan ke jenjang Sekolah Dasar di SDN 5 Penengahan pada tahun 2008 dan lulus pada tahun 2013. Selanjutnya, menempuh pendidikan di SMPN 2 Bandar Lampung dan lulus pada tahun 2016. Lalu, penulis melanjutkan pendidikan ke jenjang Sekolah Menengah Atas di SMAN 9 Bandar Lampung dan lulus pada tahun 2019. Di tahun yang sama, penulis berhasil menjadi mahasiswa di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA) Universitas Lampung melalui jalur SBMPTN.

Selama menjadi mahasiswa, penulis aktif mengikuti beberapa organisasi yang ada di kampus, seperti menjadi anggota Bidang Minat dan Bakat Himpunan Mahasiswa Jurusan Matematika (HIMATIKA) pada tahun 2020. Kemudian, penulis juga pernah menjadi anggota Koperasi Mahasiswa (Kopma) Universitas Lampung pada tahun 2021. Selain mengikuti beberapa organisasi, pada tahun 2022, penulis juga mengikuti salah satu program Merdeka Belajar Kampus Merdeka (MBKM), yaitu Studi Independen di PT. Hacktivate Teknologi Indonesia.

Sebagai bentuk aplikasi bidang ilmu di dunia kerja, penulis telah melaksanakan Kerja Praktik (KP) di Kantor Witel Lampung selama 40 hari di Bulan Januari hingga Februari tahun 2022. Lalu, sebagai bentuk aplikasi bidang ilmu kepada masyarakat, penulis telah melakukan Kuliah Kerja Nyata (KKN) selama 40 hari di Desa Way Mili, Kecamatan Gunung Pelindung, Kabupaten Lampung Timur.

KATA INSPIRASI

“Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya”

(Q.S. Al-Baqarah: 266)

*“It’s not always easy but that’s life,
be strong cause there are better days ahead”*

(Mark Lee, member of NCT)

PERSEMBAHAN

Puji dan syukur kepada Allah SWT., atas rahmat, berkah, dan ridha-Nya sehingga skripsi ini dapat terselesaikan dengan baik.

Karya sederhana ini kupersembahkan kepada:

Mama dan Papa

Sebagai tanda terima kasih atas segala perjuangan dan pengorbanan yang telah diberikan. Tak luput pula, doa, ridha, dan dukungan senantiasa kalian berikan hingga penulis bisa berada di titik ini.

Almamater Tercinta, Universitas Lampung

SANWACANA

Puji dan syukur penulis panjatkan kehadiran Allah SWT, yang telah memberikan rahmat, karunia, dan ridha-Nya sehingga penulis mampu menyelesaikan skripsi yang berjudul **“Penerapan Algoritma *K-Nearest Neighbor* pada Data Kategorik dengan Perhitungan Jarak Menggunakan *Weighted Simple Matching Coefficient*”**. Skripsi ini disusun sebagai salah satu syarat untuk mendapatkan gelar Sarjana Matematika pada Jurusan Matematika FMIPA Unila.

Dalam penyelesaian skripsi ini, penulis menyadari adanya keterbatasan pengetahuan dan kemampuan yang dimiliki, sehingga penulis mendapatkan banyak bimbingan, arahan, dan bantuan dari berbagai pihak. Oleh karena itu, pada kesempatan ini, penulis ingin menyampaikan terima kasih sebanyak-banyaknya kepada:

1. Bapak Dr. Aang Nuryaman, S.Si., M.Si. selaku Ketua Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung sekaligus dosen pembimbing utama yang selalu memberikan waktu, bimbingan, bantuan, dan saran kepada penulis dalam menyelesaikan skripsi ini.
2. Bapak Dr. Ahmad Faisol, S.Si., M.Sc. selaku dosen pembimbing kedua yang juga telah memberikan waktu, bimbingan, bantuan, dan saran kepada penulis dalam menyelesaikan skripsi ini.
3. Bapak Drs. Eri Setiawan, M.Si. selaku dosen pembahas sekaligus dosen pembimbing akademik. Terima kasih atas bimbingan, bantuan, nasihat, dan saran yang telah diberikan, baik saat seminar maupun selama penulis menjalani perkuliahan.

4. Bapak Dr. Eng. Heri Satria, S.Si., M.Si. selaku Plt. Dekan FMIPA Universitas Lampung.
5. Seluruh dosen, staff, dan karyawan Jurusan Matematika FMIPA Universitas Lampung.
6. Orang tuaku tersayang, Papa Dedi dan Mama Aike Adriani, yang selalu memberikan semangat, dukungan, serta doa kepada penulis. Terima kasih atas segala perjuangan dan pengorbanan sehingga penulis bisa berada di titik ini. Semoga penulis dapat memberikan kebahagiaan dan selalu menjadi kebanggaan Mama dan Papa.
7. Kakak Dimas, Abang Fatur, dan Raihan yang selalu memberikan dukungan dan semangat, serta senantiasa menjadi teman cerita penulis di kala penulis merasa kesulitan.
8. Om, Tante, Pakde, Bude, dan sepupu-sepupu yang selalu memberikan motivasi dan dukungan dalam setiap langkah yang penulis tempuh.
9. Sahabat-sahabat seperjuangan, Dinda, Dila, Nabilla, Grace, Putri, dan Zidny, yang selalu ada sejak awal sampai akhir perkuliahan. Terima kasih atas bantuan, semangat, dan kebersamaannya selama masa perkuliahan.
10. Sahabatku, Aldiesa, yang selalu menemani penulis dalam pengerjaan skripsi ini. Terima kasih selalu mendengarkan penulis berkeluh kesah, selalu memberikan saran dan semangat, serta selalu menyempatkan waktu, baik dalam pengerjaan skripsi, maupun saat pelaksanaan seminar yang penulis lalui.
11. Teman-teman KKN Desa Way Mili, terutama Mega, Hani, dan Faliza, yang selalu memberikan semangat dan doa kepada penulis selama pengerjaan skripsi ini.
12. Teman-teman mahasiswa Jurusan Matematika angkatan 2019.
13. Seluruh pihak yang telah membantu penulis dalam pengerjaan skripsi ini yang tidak bisa disebutkan satu persatu.

Penulis menyadari ketidaksempurnaannya skripsi ini di mana masih banyak kekurangan dalam penyajian maupun teknik penulisan. Oleh karena itu, saran dan kritik yang membangun senantiasa penulis harapkan dari seluruh pihak. Penulis berharap semoga skripsi ini dapat memberikan manfaat bagi para pembaca dan penulis sendiri.

Bandar Lampung, 11 April 2023

Penulis

Citra Puspa Tria

NPM. 1917031074

DAFTAR ISI

| | Halaman |
|---|---------|
| DAFTAR TABEL | vii |
| DAFTAR GAMBAR | ix |
| I. PENDAHULUAN | 1 |
| 1.1 Latar Belakang dan Masalah..... | 1 |
| 1.2 Tujuan Penelitian | 5 |
| 1.3 Manfaat Penelitian | 5 |
| II. TINJAUAN PUSTAKA | 6 |
| 2.1 <i>Data Mining</i> | 6 |
| 2.2 <i>Machine Learning</i> | 7 |
| 2.3 Klasifikasi | 8 |
| 2.4 Tahapan Klasifikasi | 9 |
| 2.4.1 Praproses Data | 9 |
| 2.4.2 Perancangan dan Implementasi..... | 10 |
| 2.4.3 Evaluasi Model | 10 |
| 2.5 <i>K-Nearest Neighbor</i> | 12 |
| 2.6 Laptop | 15 |
| 2.7 <i>Web Scraping</i> | 16 |
| III. METODE PENELITIAN | 17 |
| 3.1 Waktu dan Tempat Penelitian | 17 |
| 3.2 Data Penelitian | 17 |
| 3.3 Metode Penelitian | 18 |
| IV. HASIL DAN PEMBAHASAN | 20 |
| 4.1 Eksplorasi Data | 20 |

| | | |
|-----------|---|-----------|
| 4.2 | Praproses Data | 22 |
| 4.2.1 | Reduksi Data | 22 |
| 4.2.2 | Pembersihan Data | 22 |
| 4.2.3 | Transformasi Data..... | 24 |
| 4.3 | Algoritma <i>K-Nearest Neighbor</i> | 26 |
| 4.3.1 | Pembagian Data | 26 |
| 4.3.2 | Perhitungan Bobot Fitur..... | 26 |
| 4.3.3 | Perhitungan WSMC | 39 |
| 4.3.4 | Penentuan Kelas | 43 |
| 4.4 | Evaluasi Model | 44 |
| V. | KESIMPULAN | 50 |
| 5.1 | Kesimpulan | 50 |
| 5.2 | Saran | 50 |
| | DAFTAR PUSTAKA | 52 |
| | LAMPIRAN | 56 |

DAFTAR TABEL

| Tabel | Halaman |
|--|---------|
| 1. <i>Confusion matrix</i> | 11 |
| 2. Variabel data harga jual laptop | 18 |
| 3. Statistika deskriptif variabel harga | 20 |
| 4. Hasil diskritisasi data pada variabel harga..... | 22 |
| 5. Jumlah data hilang pada data harga jual laptop | 23 |
| 6. Hasil <i>encoding</i> label untuk setiap variabel | 24 |
| 7. Distribusi frekuensi tiap kategori untuk fitur merek | 27 |
| 8. Nilai $p(m s_d)$ untuk fitur merek | 28 |
| 9. Distribusi frekuensi kategori 0-14 untuk fitur prosesor..... | 28 |
| 10. Distribusi frekuensi kategori 15-28 untuk fitur prosesor..... | 28 |
| 11. Nilai $p(m s_d)$ untuk kategori 0-9 di fitur prosesor | 29 |
| 12. Nilai $p(m s_d)$ untuk kategori 10-19 di fitur prosesor | 29 |
| 13. Nilai $p(m s_d)$ untuk kategori 20-28 di fitur prosesor | 29 |
| 14. Distribusi frekuensi tiap kategori untuk fitur RAM | 29 |
| 15. Nilai $p(m s_d)$ untuk fitur RAM | 30 |
| 16. Distribusi frekuensi tiap kategori untuk fitur penyimpanan | 30 |

| | | |
|-----|---|----|
| 17. | Nilai $p(m s_d)$ untuk fitur penyimpanan..... | 30 |
| 18. | <i>Global entropy</i> setiap kategori pada masing-masing fitur..... | 32 |
| 19. | Peluang munculnya setiap kategori pada masing-masing fitur | 35 |
| 20. | Bobot fitur data latih sebesar 80% | 38 |
| 21. | Bobot fitur data latih sebesar 90% | 38 |
| 22. | Data latih..... | 39 |
| 23. | Data uji..... | 40 |
| 24. | Hasil SMC antara data latih dan sampel data uji..... | 40 |
| 25. | Hasil WSMC..... | 42 |
| 26. | Tiga belas data dengan nilai WSMC terkecil | 43 |
| 27. | <i>Confusion matrix</i> hasil klasifikasi data uji berdasarkan 80% data latih dan jumlah k sebanyak 13 | 44 |
| 28. | Nilai akurasi..... | 45 |
| 29. | Nilai <i>recall</i> | 47 |
| 30. | Nilai presisi..... | 48 |
| 31. | Nilai <i>f1-score</i> | 49 |

DAFTAR GAMBAR

| Gambar | Halaman |
|--|---------|
| 1. Diagram alir algoritma KNN dengan ukuran jarak WSMC | 15 |
| 2. Diagram alir proses klasifikasi dengan algoritma KNN..... | 19 |
| 3. Histogram variabel harga..... | 21 |
| 4. Nilai akurasi hasil klasifikasi untuk tiap k | 46 |

I. PENDAHULUAN

1.1 Latar Belakang dan Masalah

Teknologi yang terus mengalami pembaharuan telah membantu manusia dalam menyelesaikan banyak hal. Salah satu bentuk teknologi yang banyak diaplikasikan dalam aspek kehidupan manusia adalah *artificial intelligence* atau kecerdasan buatan. Kecerdasan buatan belajar untuk mengidentifikasi pola dalam dataset yang besar sehingga dapat melaksanakan tugas yang kompleks.

Statistika sebagai salah satu cabang ilmu yang banyak digunakan di kehidupan sehari-hari juga menerapkan konsep kecerdasan buatan ini. Salah satu metode statistika yang menggunakan kecerdasan buatan dalam prosesnya adalah *data mining*. *Data mining* merupakan salah satu tahapan dalam *knowledge discovery in database* (KDD) (Fitriani & Yustanti, 2022). Proses *data mining* melibatkan metode statistika, matematika, kecerdasan buatan, dan *machine learning* untuk memperoleh informasi dari dataset yang besar (Turban, dkk., 2005). Informasi yang diperoleh dari *data mining* tersebut dapat dijadikan alat pendukung keputusan dalam penyelesaian suatu masalah.

Pembelajaran pola dalam proses *data mining* dilakukan oleh *machine learning*. *Machine learning* sebagai bentuk kecerdasan buatan, akan melatih komputer untuk mempelajari suatu dataset sehingga dapat menghasilkan suatu keputusan atau prediksi tanpa harus menggunakan kode program secara berulang (Lombu, dkk., 2022). *Machine learning* sendiri terbagi lagi menjadi 2, yaitu *supervised learning* (pembelajaran terawasi) dan *unsupervised learning* (pembelajaran tak terawasi). *Unsupervised learning* bekerja dalam data masukan yang tidak memiliki label sehingga pembelajaran ini membiarkan komputer untuk menyelesaikan masalah dengan menelusuri karakteristik data itu sendiri (Lashari, dkk., 2018). Sedangkan, *supervised learning* bekerja dalam data yang memiliki label sehingga keluaran yang diharapkan adalah komputer dapat memberikan label dari data yang belum memiliki label (Retnoningsih & Pramudita, 2020). Salah satu bentuk *supervised learning* adalah klasifikasi.

Klasifikasi bertujuan untuk memprediksi kelas dari suatu objek yang belum diketahui kelasnya berdasarkan proses perhitungan dari data lampau dengan menggunakan suatu algoritma (Indrayanti, dkk., 2017). Akan tetapi, klasifikasi sebagai salah satu bentuk *supervised learning*, memerlukan data latih untuk menemukan pola klasifikasi. Selanjutnya, pola klasifikasi yang telah dibangun dengan menggunakan algoritma yang ada akan diimplementasikan pada data uji (Wafiyah, dkk., 2017). Salah satu algoritma klasifikasi yang banyak digunakan adalah *k-nearest neighbor* (KNN).

Algoritma KNN mengklasifikasikan objek berdasarkan kelas mayoritas dari k objek terdekat di sekitarnya (Afandie, dkk., 2014). Dengan algoritmanya yang sederhana, metode ini tepat digunakan untuk dataset yang berukuran kecil (Suguna & Thanushkodi, 2010). Namun, algoritma ini sangat bergantung dengan nilai k sehingga pemilihan nilai k sangat mempengaruhi hasil klasifikasi (Al Karomi, 2015). Pada umumnya, untuk melihat k objek terdekat dapat dilakukan dengan melihat jarak antar objek. Ukuran jarak yang biasa digunakan adalah jarak *Euclidean*, seperti

penelitian yang dilakukan oleh Nurjanah, dkk. (2020). Pada penelitian tersebut, peneliti menerapkan algoritma KNN dengan menggunakan fitur berupa data numerik (Nurjanah, dkk., 2020).

Namun, pengukuran jarak dengan perhitungan tersebut tidak sesuai apabila data yang digunakan merupakan data kategorik (Larose, 2005). Leidiyana (2013) melakukan penelitian mengenai klasifikasi dengan algoritma KNN pada data kategorik. Leidiyana menghitung kedekatan antar objek berdasarkan bobot fitur dan kesamaan antara data latih dan data uji. Akurasi yang dihasilkan dalam penelitian ini adalah 81.46% (Leidiyana, 2013). Selain itu, terdapat ukuran jarak yang umum digunakan untuk data kategorik, yaitu *simple matching coefficient* (SMC). Lalu, dengan memberikan bobot pada setiap fitur dalam perhitungan memunculkan metode *weighted simple matching coefficient* (WSMC). Perhitungan bobot didasari atas informasi entropi. Salah satu penelitian yang menerapkan algoritma KNN dengan ukuran jarak antar objek menggunakan WSMC adalah penelitian yang dilakukan oleh Wijaya, dkk. (2017). Peneliti tersebut menerapkan algoritma *nearest neighbor* untuk 6 dataset berbeda dengan hasil akurasi berada di atas 72% untuk setiap datasetnya (Wijaya, dkk., 2017).

Klasifikasi, terutama KNN, banyak digunakan dalam berbagai bidang. Beberapa di antaranya adalah kesehatan, pendidikan, teknik bangunan, pertanian, dan lain sebagainya. Salah satu penelitian bidang kesehatan yang menerapkan algoritma KNN adalah penelitian yang dilakukan oleh Naufal, dkk. (2020). Mereka membandingkan algoritma *support vector machine* (SVM) dan KNN untuk mendeteksi kanker pada data *microarray*. Akurasi yang dihasilkan dari algoritma KNN adalah 100% untuk data kanker paru-paru, 57.4% untuk data kanker payudara, dan 85.5% untuk kanker usus besar (Naufal, dkk., 2020). Selain itu, terdapat penelitian bidang pendidikan yang menerapkan algoritma KNN yang dilakukan oleh Purwaningsih & Nurelasari (2021). Mereka mengklasifikasikan tingkat kelulusan siswa pada salah satu sekolah swasta di Bekasi dari tahun 2013 hingga 2014 dengan akurasi hasil klasifikasi yang

diperoleh sebesar 96.49% (Purwaningsih & Nurelasari, 2021). Selanjutnya, penelitian di bidang pertanian, di mana Liantoni (2015) mengklasifikasikan daun dengan perbaikan fitur citra. Nilai akurasi dari hasil penelitian ini adalah 86.67% (Liantoni, 2015).

Akan tetapi, tidak hanya pada bidang yang dijelaskan sebelumnya, klasifikasi KNN juga dapat diterapkan untuk pemecahan masalah sehari-hari. Contohnya adalah pengambilan keputusan dalam membeli barang, salah satunya laptop. Laptop merupakan salah satu bentuk perkembangan teknologi yang saat ini sangat dibutuhkan. Bentuknya yang ringan dan fleksibel sehingga pengguna dapat membawa dan menggunakannya di manapun. Perkembangan tipe dan jenis laptop yang diikuti oleh semakin beragamnya fitur yang ditawarkan menyebabkan orang kesulitan untuk memilih laptop yang sesuai dengan kebutuhan. Dalam penelitian yang dilakukan oleh Sunarsa & Handayani (2016), mereka meneliti kriteria utama dalam pemilihan laptop dengan menggunakan metode *Analytical Hierarchy Process* (AHP). Analisis yang mereka lakukan menunjukkan hasil bahwa kriteria paling penting dalam menentukan pilihan laptop oleh PT.INDOTEKNO adalah merek, *processor*, VGA, RAM, kapasitas HDD, keunggulan, dan harga (Sunarsa & Handayani, 2016). Selain itu, penelitian dengan metode yang sama juga dilakukan oleh Casym & Oktiara (2020). Berdasarkan hasil analisis, kriteria terbesar dalam pemilihan laptop bagi mahasiswa di Jakarta Utara adalah harga, *processor*, dan RAM (Casym & Oktiara, 2020). Oleh karena itu, pada penelitian ini akan dilakukan perancangan pola klasifikasi harga laptop berdasarkan fitur laptop untuk membantu pembeli memutuskan pembelian laptop dengan menggunakan algoritma KNN.

1.2 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah:

1. Merancang pola klasifikasi untuk menentukan harga laptop berdasarkan fitur-fitur yang ada dengan menggunakan metode KNN.
2. Melihat performa klasifikasi dengan melihat nilai akurasi, *recall*, presisi dan *F1-score*.

1.3 Manfaat Penelitian

Adapun manfaat yang dapat diambil dari penelitian ini adalah:

1. Sebagai rujukan pengembangan ilmu pengetahuan di bidang matematika bagi peneliti dalam membuat klasifikasi menggunakan metode KNN.
2. Sebagai bahan pertimbangan dalam mengambil keputusan pembelian laptop atau tidak berdasarkan harga yang ditawarkan.

II. TINJAUAN PUSTAKA

2.1 *Data Mining*

Suatu proses menemukan pola dataset atau kasus lama dengan cara menganalisa menggunakan metode statistika dan matematika disebut dengan *data mining* (Al Karomi, 2015). Proses *data mining* tersebut nantinya dapat memberikan informasi terkait data yang sebelumnya yang mungkin saja tidak berguna. Dalam pengaplikasiannya, data yang digunakan dapat berbentuk tabulasi, berbasis teks, maupun multimedia (citra) (Nurjanah, dkk., 2020).

Data mining kerap kali dianggap memiliki makna yang sama dengan *Knowledge Discovery in Database* (KDD). Padahal, *data mining* merupakan salah satu tahapan dari proses *Knowledge Discovery in Database* (KDD) (Fitriani & Yustanti, 2022). Adapun proses keseluruhan dari KDD adalah *data selection*, *data preprocessing*, *data transformation*, *data mining*, dan *evaluation*. *Data mining* dalam KDD berguna untuk mengenali pola dari data (Krishnaiah, dkk., 2013).

Berdasarkan tugasnya, *data mining* dikelompokkan menjadi deskriptif dan prediktif. *Data mining* deskriptif menghasilkan karakteristik dan deskripsi mengenai dataset, sedangkan *data mining* prediktif menghasilkan prediksi berdasarkan pola dari data

yang diketahui. Beberapa metode *data mining* yang banyak digunakan antara lain klasifikasi, klusterisasi, dan asosiasi (Lashari, dkk., 2018).

2.2 Machine Learning

Machine learning atau pembelajaran mesin merupakan salah satu bentuk dari *artificial intelligence* (kecerdasan buatan) yang melatih komputer untuk belajar berdasarkan suatu data sehingga dapat menghasilkan suatu keputusan atau prediksi tanpa harus menggunakan kode program secara berulang. Istilah ini pertama kali dipopulerkan oleh seorang ilmuwan komputer yang bernama Arthur Samuel pada tahun 1959 (Lombu, dkk., 2022). Berdasarkan teknik pembelajarannya, *machine learning* terbagi menjadi 2, yaitu:

a. Supervised Learning (Pembelajaran Terawasi)

Pembelajaran dengan teknik ini menggunakan masukan data yang telah memiliki label sehingga komputer dapat belajar mengenali label masukan berdasarkan fitur yang ada untuk selanjutnya dapat menghasilkan keluaran berupa label dari data baru yang belum memiliki label (Retnoningsih & Pramudita, 2020). Tugas dalam pembelajaran terawasi dapat berupa klasifikasi dan regresi. Label yang digunakan dalam klasifikasi berupa data nominal, sedangkan dalam regresi berupa data numerik (Lashari, dkk., 2018).

b. Unsupervised Learning (Pembelajaran Tak Terawasi)

Pembelajaran tak terawasi berfungsi dalam kondisi keluaran yang tidak diketahui sehingga pembelajaran ini membiarkan komputer untuk menyelesaikan masalah dengan menelusuri karakteristik data itu sendiri. Pembelajaran ini bekerja dalam data yang tidak memiliki label (Lashari, dkk., 2018). Salah satu tugas dalam pembelajaran tak terawasi adalah klusterisasi. Hasil klusterisasi berupa label yang mengelompokkan data berdasarkan kesamaan karakteristik data.

2.3 Klasifikasi

Klasifikasi merupakan salah satu bentuk pembelajaran terawasi yang bertujuan untuk memprediksi kelas data suatu objek yang belum diketahui labelnya berdasarkan proses perhitungan dari data lampau dengan menggunakan suatu algoritma (Indrayanti, dkk., 2017). Dalam klasifikasi terdapat beberapa komponen penting, yaitu (Gorunescu, 2011):

a. Kelas

Kelas merupakan data kategorik yang merepresentasikan label hasil klasifikasi. Label atau juga dapat disebut data target merupakan variabel dependen dalam model. Berdasarkan kelasnya, klasifikasi terbagi menjadi 2, yaitu klasifikasi biner dan multi-kelas. Klasifikasi biner terjadi ketika kelas dalam label berjumlah 2, sedangkan klasifikasi multi-kelas terjadi ketika kelas dalam label berjumlah lebih dari 2 (Kesavaraj & Sukumaran, 2013).

b. Prediktor

Prediktor atau dapat disebut fitur adalah variabel independen yang direpresentasikan oleh atribut data yang nantinya akan digunakan untuk pengklasifikasian. Banyaknya atribut mempengaruhi performa suatu algoritma (Prasetyo, 2012).

c. Dataset Pembelajaran

Data pembelajaran berisi nilai dari 2 komponen di atas yang digunakan untuk proses pembelajaran sehingga dengan didasari oleh prediktor, kelas dapat ditentukan oleh program.

d. Dataset Pengujian

Dataset pengujian berisi data yang tidak diketahui labelnya sehingga akan ditentukan kelasnya berdasarkan pola yang dihasilkan pada proses pembelajaran. Dari dataset pengujian ini juga dapat dilihat nilai akurasi klasifikasi.

2.4 Tahapan Klasifikasi

2.4.1 Praproses Data

Banyak peneliti menyatakan bahwa praproses data merupakan hal yang penting dan krusial dalam proses KDD. Hal itu menandakan bahwa praproses data juga penting dalam proses klasifikasi sebagai salah satu bentuk KDD. Tujuan dari praproses data adalah mengatasi masalah dalam data yang belum diproses, termasuk data tidak konsisten, *noise*, data hilang, pencilan, dimensi tinggi, dan data tidak seimbang (Benhar, dkk., 2020). Terdapat 5 bentuk praproses data, yaitu sebagai berikut:

1. Pembersihan Data

Proses mengatasi permasalahan data hilang, mencari dan menghapus pencilan atau *noise*, serta mengatasi ketak-konsistenan data merupakan proses pembersihan data. Data hilang merupakan masalah yang paling sering muncul dalam dataset. Dalam proses klasifikasi, data hilang, baik yang ada di data latih maupun data uji, dapat mempengaruhi akurasi prediksi (Zhang, dkk., 2012). Namun, masalah ini dapat diatasi dengan beberapa cara, di antaranya adalah dengan menghapus data yang hilang atau dengan menambahkan nilai pada *cell* data hilang tersebut.

2. Reduksi Data

Reduksi data merupakan proses mereduksi data fitur dalam dataset. Beberapa bentuk proses mereduksi data adalah pemilihan fitur dan transformasi data fitur kontinu ke jenis data nominal.

3. Transformasi Data

Transformasi data meliputi *encoding* data kategorik dan standarisasi data untuk mengurangi perbedaan yang besar dalam skala antar atribut.

4. Integrasi Data

Integrasi data merupakan proses mengumpulkan data dari berbagai sumber dan menyatukannya ke dalam dataset yang koheren dan logis.

5. Penyeimbangan Data

Ketidakseimbangan data merupakan suatu masalah di mana jumlah data antar kelas berbeda jauh. Permasalahan tersebut dapat teratasi dengan melakukan penyeimbangan data.

2.4.2 Perancangan dan Implementasi

Proses klasifikasi dimulai dengan melatih data untuk menemukan pola klasifikasi. Oleh karena terdapat proses pelatihan, maka dilakukan pembagian data menjadi data latih dan data uji. Pembagian data dilakukan secara acak dengan proporsi yang sama untuk masing-masing kelas (Hastuti, 2012). Proses pelatihan data dilakukan dengan menerapkan algoritma-algoritma klasifikasi yang ada, seperti *decision tree*, *support vector machine*, dan *k-nearest neighbor*. Selanjutnya, pola klasifikasi yang telah dibangun pada proses perancangan akan diimplementasikan pada data uji. Proses ini dinamakan proses implementasi (Wafiyah, dkk., 2017).

2.4.3 Evaluasi Model

Dalam klasifikasi, penting untuk melihat ketepatan klasifikasi. Perhitungan ketepatan klasifikasi dapat dengan melihat *confusion matrix*. *Confusion matrix* berisi perbandingan antara kelas hasil klasifikasi dengan kelas sebenarnya pada data uji (Indrayanti, dkk., 2017). *Confusion matrix* berukuran $N \times N$ yang mana bagian kolom merepresentasikan kelas prediksi dan bagian baris merepresentasikan kelas sebenarnya (Markoulidakis, dkk., 2021). Dalam klasifikasi biner, N berjumlah 2 yang merepresentasikan 2 kelas klasifikasi.

Dalam *confusion matrix*, terdapat beberapa istilah yang merepresentasikan kondisi perbandingan antara kelas hasil prediksi dan yang sebenarnya, seperti sebagai berikut:

- a. Kondisi ketika data dengan kelas prediksi positif dan kelas yang sebenarnya juga positif disebut dengan *True Positive* (TP).
- b. Kondisi ketika kelas hasil prediksi dan yang sebenarnya sama-sama negatif, dinamakan dengan *True Negative* (TN).
- c. Apabila kelas hasil prediksi adalah positif namun kelas yang sebenarnya adalah negatif, maka keadaan ini disebut dengan *False Positive* (FP).
- d. Apabila kelas hasil prediksi adalah negatif namun kelas yang sebenarnya adalah positif, maka keadaan ini disebut dengan *False Negative* (FN)

Tabel 1. *Confusion matrix*

| Klasifikasi | | Kelas Prediksi | |
|------------------|-----------------|------------------------------|------------------------------|
| | | <i>Positive</i> | <i>Negative</i> |
| Kelas Sebenarnya | <i>Positive</i> | TP (<i>True Positive</i>) | FN (<i>False Negative</i>) |
| | <i>Negative</i> | FP (<i>False Positive</i>) | TN (<i>True Negative</i>) |

Berdasarkan *confusion matrix*, dapat dihitung beberapa perhitungan metrik untuk melihat performa suatu algoritma klasifikasi, seperti sebagai berikut:

1) Akurasi

Akurasi merupakan persentase kemungkinan hasil klasifikasi adalah benar. Nilai akurasi dapat dihitung dengan rumus sebagai berikut:

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2) Recall

Recall merupakan proporsi jumlah prediksi kelas positif yang benar dari keseluruhan kelas positif yang sebenarnya. Perhitungan *recall* berguna untuk melihat sensitivitas

pola klasifikasi dalam memprediksi data ke dalam kelas positif (Grandini, dkk., 2020). *Recall* dapat dihitung dengan rumus sebagai berikut:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

3) Presisi

Presisi merupakan persentase proporsi jumlah prediksi kelas positif yang benar dari keseluruhan hasil yang memprediksi kelas positif tersebut. Presisi juga dapat disebut sebagai derajat reliabilitas klasifikasi ketika sebuah data diklasifikasikan ke dalam kelas positif (Grandini, dkk., 2020). Perhitungan presisi dapat dilakukan dengan menggunakan rumus sebagai berikut:

$$Presisi = \frac{TP}{TP + FP} \quad (3)$$

4) *F1-Score*

F1-score mengevaluasi efektivitas pola klasifikasi dengan menghitung rata-rata harmonik *recall* dan presisi. Berikut rumus untuk mencari nilai *f1-score*:

$$F_1Score = 2 \times \left(\frac{Recall \times Presisi}{Recall + Presisi} \right) \quad (4)$$

2.5 *K-Nearest Neighbor*

K-nearest neighbor (KNN) merupakan salah satu metode yang menerapkan algoritma *supervised learning* dengan tujuan mengklasifikasikan objek berdasarkan *k* objek terdekat di sekitarnya. Hasil klasifikasi didasari atas kelas mayoritas dari *k* objek terdekat (Afandie, dkk., 2014). Oleh karena itu, pemilihan nilai *k* sangatlah penting dalam melakukan algoritma KNN ini.

Metode KNN merupakan salah satu metode yang sederhana dan paling banyak digunakan dalam klasifikasi. Dengan algoritmanya yang sederhana, metode ini tepat digunakan untuk dataset yang berukuran kecil (Suguna & Thanushkodi, 2010). Namun, seperti yang telah diketahui, metode ini sangat bergantung dengan nilai k karena dapat mempengaruhi hasil klasifikasi. Nilai k yang besar dapat mengakibatkan hasil klasifikasi menjadi kabur karena batasan mungkin akan menjadi lebih besar (Al Karomi, 2015).

Perhitungan jarak antar objek dapat dilakukan dengan menggunakan ukuran jarak, seperti jarak *Euclidean* dan jarak *Manhattan*. Namun, pengukuran jarak dengan ukuran jarak tersebut tidak sesuai apabila data yang digunakan berupa data kategorik (Larose, 2005). Penggunaan ukuran jarak untuk data kategorik yang cukup umum digunakan adalah *simple matching coefficient* yang didefinisikan sebagai berikut:

$$SMC(\mathbf{x}_i, \mathbf{x}_j) = \sum_{d=1}^D I(x_{id} \neq x_{jd}) \quad (5)$$

dengan $I(true) = 1$ dan $I(false) = 0$. Namun, SMC mengasumsikan seluruh fitur memiliki bobot yang sama dalam prediksi kelas. Oleh karena itu, dikembangkan metode *weighted simple matching coefficient* (WSMC) yang memberikan bobot pada setiap fitur dalam perhitungan (Chen & Guoi, 2015).

Pembobotan yang digunakan dalam WSMC dibagi menjadi dua, yaitu pembobotan global dan lokal. Pada metode global, bobot dihitung untuk tiap atribut, sedangkan bobot pada metode lokal dihitung untuk setiap kelas pada atribut yang digunakan. Berikut rumus perhitungan jarak dengan metode $WSMC_{global}$:

$$WSMC_{global}(\mathbf{x}_i, \mathbf{x}_j, \bar{\omega}) = \sum_{d=1}^D \omega_d \times I(x_{id} \neq x_{jd}) \quad (6)$$

dengan ω_d merupakan *global weight* yang dapat dihitung berdasarkan informasi entropi dengan rumus sebagai berikut:

$$\omega_d^{GE} = e^{-\frac{1}{\log_2 M} \sum_{s_d \in S_d} p(s_d) \times GE(s_d)} \quad (7)$$

Selanjutnya, $GE(s_d)$ atau *global entropy* dan $p(m|s_d)$ atau peluang suatu kategori muncul di setiap kelas didefinisikan sebagai berikut:

$$GE(s_d) = - \sum_{m=1}^M p(m|s_d) \times \log_2 p(m|s_d) \quad (8)$$

$$p(m|s_d) = \frac{\sum_{(x,y) \in c_m} I(x_d = s_d)}{\sum_{(x,y) \in tr} I(x_d = s_d)} \quad (9)$$

Keterangan:

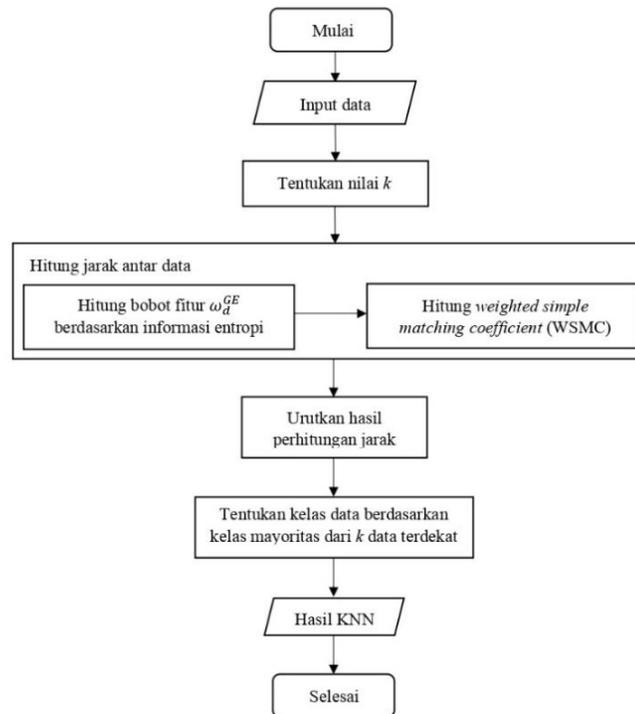
s_d = kategori ke- d dalam sebuah fitur

M = jumlah kelas yang ada pada dataset

tr = data latih

c_m = kelas ke- m

Berikut diagram alir algoritma KNN dengan perhitungan ukuran jarak menggunakan WSMC.



Gambar 1. Diagram alir algoritma KNN dengan ukuran jarak WSMC

2.6 Laptop

Laptop merupakan salah satu bentuk perkembangan teknologi yang mana sering disebut juga sebagai komputer jinjing. Bentuk komputer yang besar dan berat menyebabkan pengguna sulit membawa komputer kemana-mana sehingga para ilmuwan termotivasi untuk menciptakan komputer dalam bentuk yang lebih minimalis. Oleh karena itu, pengguna dapat membawa dan menggunakannya di manapun.

Alan Kay merupakan ilmuwan komputer dari Amerika yang pertama kali mencetuskan gagasan pembuatan laptop. Gagasan beliau dicetuskan sejak tahun 1970-an dan diberi nama “*Osborne I*”. Kemudian, pada bulan April 1981, “*Osborne I*” dirilis oleh *Osborne Computer Corporation*. Namun, awalnya laptop buatan Alan Kay tersebut masih berukuran besar dan membutuhkan sambungan listrik. Lalu, seiring dengan perkembangan jaman, laptop sudah berukuran 10-17 inci dan tidak membutuhkan sambungan listrik secara terus menerus (Thandung, dkk., 2013). Tidak hanya itu, perkembangan jaman juga menjadikan laptop memiliki fitur-fitur yang berguna untuk memaksimalkan performa penggunaannya, seperti OS (*Operating System*), *processor*, RAM, kapasitas penyimpanan, VGA (*Video Graphics Array*), dan lain sebagainya. Semakin tinggi versi fitur-fitur tersebut maka akan semakin cepat pemrosesan kegiatan laptop (Ayyubi, dkk., 2022).

2.7 Web Scraping

Web scraping adalah proses pengambilan informasi dari situs *web* secara otomatis tanpa harus menyalinnya secara manual (Ayani, dkk., 2019). *Web scraping* atau yang dikenal dengan ekstrasi *web* juga dapat diartikan sebagai metode pengambilan data dari situs *web* lalu menyimpannya dalam *file* atau database untuk keperluan analisis data. Proses ekstrasi data dari situs *web* ini dimulai dengan menentukan situs *web* yang akan diekstrak datanya. Lalu, mengekstrasi data dari situs *web* tersebut sesuai dengan kebutuhan (Zhao, 2017). Salah satu alat untuk melakukan *web scraping* adalah dengan menggunakan pemrograman *python* (Mitchell, 2018). Selain itu, mesin pencarian (*browser*) juga menyediakan aplikasi ekstensi untuk melakukan proses ekstrasi data, seperti *Web Scraper* dan *Instant Data Scraper* (Djufri, 2020).

III. METODE PENELITIAN

3.1 Waktu dan Tempat Penelitian

Penelitian ini dilakukan pada semester ganjil tahun akademik 2022/2023 bertempat di Jurusan Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lampung.

3.2 Data Penelitian

Data yang digunakan merupakan data sekunder, yaitu data harga jual laptop yang diperoleh dari laman Tokopedia dengan jumlah data sebanyak 1072 data. Fitur yang digunakan adalah merek laptop, tipe *processor*, besar RAM, dan kapasitas penyimpanan laptop. Variabel target dalam penelitian ini adalah harga dengan kelas berjumlah 2. Keempat fitur merupakan data kategorik dengan jumlah kategori sebagai berikut.

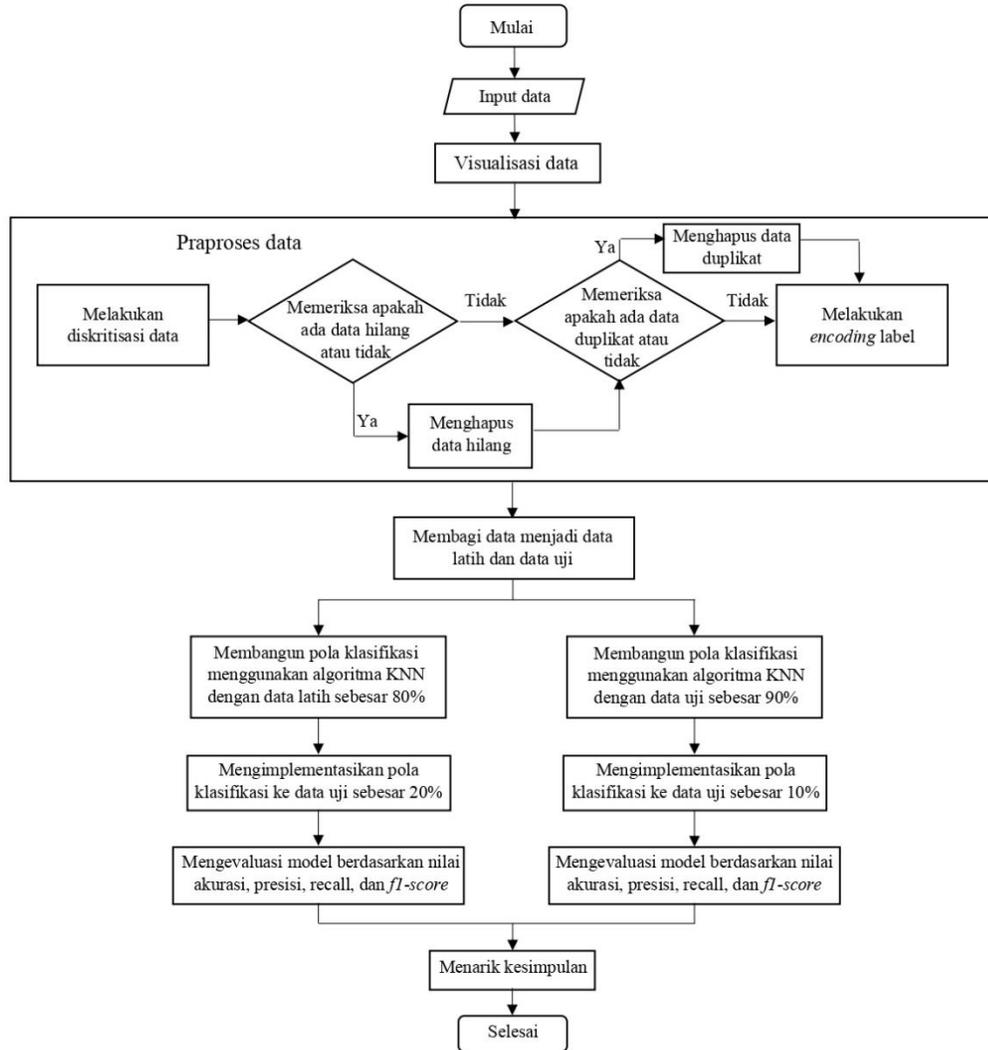
Tabel 2. Variabel data harga jual laptop

| No. | Variabel | Jenis Data | Jumlah kategori |
|-----|-----------------------|----------------|-----------------|
| 1. | Merek | Data kategorik | 4 |
| 2. | Prosesor | Data kategorik | 29 |
| 3. | RAM | Data kategorik | 4 |
| 4. | Kapasitas Penyimpanan | Data kategorik | 6 |

3.3 Metode Penelitian

Penelitian ini menggunakan algoritma *k-nearest neighbor* untuk melihat klasifikasi harga laptop berdasarkan 4 fitur yang digunakan dengan bantuan bahasa pemrograman *Python*. Adapun langkah-langkah penelitian ini adalah sebagai berikut.

1. Memasukkan data harga laptop yang diperoleh dari situs *web* Tokopedia ke dalam bahasa pemrograman *Python* menggunakan *Jupyter Notebook*.
2. Melakukan analisis deskriptif melalui visualisasi data berupa histogram.
3. Melakukan *preprocessing* data, di antaranya:
 - a. Melakukan diskritisasi pada variabel harga.
 - b. Memeriksa ada atau tidaknya data hilang dan juga data duplikat.
 - c. Melakukan *encoding* pada data kategorik.
4. Membagi data dalam 2 skema, yaitu:
 - a. Sebanyak 80% untuk data latih dan 20% untuk data uji
 - b. Sebanyak 90% untuk data latih dan 10% untuk data uji.
5. Membangun pola klasifikasi dengan algoritma *k-nearest neighbor* dengan menggunakan data latih.
6. Mengimplementasikan pola klasifikasi ke data uji.
7. Mengevaluasi model berdasarkan nilai akurasi, presisi, *recall*, dan *f1-score*.



Gambar 2. Diagram Alir Proses Klasifikasi dengan Algoritma KNN

V. KESIMPULAN

5.1 Kesimpulan

Berdasarkan hasil dan pembahasan mengenai algoritma KNN pada data kategorik dengan menggunakan WSMC sebagai ukuran kedekatan antar objek dan juga membandingkan dua skema pembagian data, diperoleh beberapa kesimpulan sebagai berikut:

1. Jumlah k atau tetangga dalam penerapan algoritma KNN yang menghasilkan akurasi terbaik adalah sebanyak 13 untuk data latih 80% dan 17 untuk data latih 90%.
2. Penerapan algoritma KNN dengan proporsi data latih 90% memberikan hasil klasifikasi yang lebih baik dibandingkan dengan proporsi data latih 80% di mana nilai akurasi yang dihasilkan sebesar 86.96%, *recall* sebesar 50%, presisi sebesar 66.67%, dan *f1-score* sebesar 57.14%.

5.2 Saran

Adapun saran terkait dengan penelitian ini, bisa dilanjutkan untuk menyelesaikan masalah, seperti sebagai berikut:

1. Salah satu penanganan masalah data hilang pada data numerik adalah dengan melakukan imputasi menggunakan nilai mean dari data. Namun, dalam

penelitian ini, data yang digunakan adalah data kategorik di mana data kategorik tidak memiliki mean. Oleh karena itu, diperlukan metode untuk penanganan masalah data hilang pada data kategorik.

2. Masih terbatasnya data yang digunakan, sehingga dapat dilakukan penelitian dengan metode yang sama untuk data yang telah diperbarui, seperti data dengan penambahan fitur, yaitu VGA atau spesifikasi laptop lainnya.
3. Melakukan perbandingan model KNN dengan penggunaan WSMC global dan lokal dengan informasi entropi dan gini, sehingga dapat memperoleh model yang lebih optimal.

DAFTAR PUSTAKA

- Afandie, M.N., Cholissodin, I., & Supianto, A.A. 2014. Implementasi Metode K-Nearest Neighbor Untuk Pendukung Keputusan Pemilihan Menu Makanan Sehat dan Bergizi. *DORO: Repository Jurnal Mahasiswa FILKOM Universitas Brawijaya*. **3**(1).
- Al Karomi, M.A. 2015. Optimasi Parameter K pada Algoritma KNN untuk Klasifikasi Heregistrasi Mahasiswa. *Jurnal IC-Tech*.
- Ayani, D.D., Pratiwi, H.S., & Muhandi, H. 2019. Implementasi Web Scraping untuk Pengambilan Data pada Situs Marketplace. *Jurnal Sistem Dan Teknologi Informasi (JUSTIN)*. **7**(4): 257-62.
- Ayyubi, S.Al, Amalia, I.S., & Arifiyanti, A.A. 2022. Komparasi Klasifikasi RAM dengan Algoritma Decision Tree dan KNN. *Procedia of Engineering and Life Science*. **2**(2).
- Benhar, H., Idri, A., & Fernandez-Aleman, J.L. 2020. Data Preprocessing for Heart Disease Classification: A Systematic Literature Review. *Computer Methods and Programs in Biomedicine*.
- Casym, J.E.S., & Oktiara, D.N. 2020. Aplikasi Analytical Hierarchy Process dalam Mengidentifikasi Preferensi Laptop Bagi Mahasiswa. *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*. **1**(1): 636-40.
- Chen, L., & Guoi, G. 2015. Nearest Neighbor Classification of Categorical Data by Attributes Weighting. *Expert System with Applications*. **42**(6): 3142-9.

- Djufri, M. 2020. Penerapan Teknik Web Scraping untuk Penggalan Potensi Pajak (Studi Kasus pada Online Market Place Tokopedia, Shopee, dan Bukalapak). *Jurnal BPPK*. **13**(2): 65-75.
- Fitriani, E.E., & Yustanti, W. 2022. Perbandingan Kinerja Metode Problem Transformation-KNN dan Algorithm Adaptation-KNN pada Klasifikasi Multi-Label Pertanyaan Kotakode. *Journal of Emerging Information Systems and Business Intelligence*. **3**(3): 122–9.
- Gorunescu, F. 2011. *Data Mining: Concepts, Models and Techniques*. Springer Science & Business Media.
- Grandini, M., Bagli, E., & Visani, G. 2020. Metrics for Multi-Class Classification: An Overview. *ArXiv Preprint ArXiv:2008.05756*.
- Hastuti, K. 2012. Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Mahasiswa Non Aktif. *Seminar Nasional Teknologi Informasi & Komunikasi Terapan*. **2**(1).
- Indrayanti, Sugianti, D., & Al Karomi, M.A. 2017. Optimasi Parameter K Pada Algoritma K-Nearest Neighbor Untuk Klasifikasi Penyakit Diabetes Mellitus. *Prosiding SNATIF Ke-4 Tahun 2017*. 153–60.
- Kesavaraj, G., & Sukumaran, S. 2013. A Study on Classification Techniques in Data Mining. *4th International Conference on Computing, Communications and Networking Technologies (ICCCNT)*.
- Krishnaiah, V., Narsimha, G., & Chandra, N.S. 2013. Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques. *International Journal of Computer Science and Information Technologies*. **4**(1): 39–45.
- Larose, D.T. 2005. *Discovering Knowledge in Data: An Introduction to Data Mining*. Jon Wiley & Sons, Inc., New Jersey.
- Lashari, S.A., Ibrahim, R., Senan, N., & Taujuddin, N.S.A.M. 2018. Application of Data Mining Techniques for Medical Data Classification: A Review. *MATEC Web of Conferences*.

- Leidiyana, H. 2013. Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Resiko Kredit Kepemilikan Kendaraan Bermotor. *Jurnal Penelitian Ilmu Komputer, System Embedded & Logic*. **1**(1): 65–76.
- Liantoni, F. 2015. Klasifikasi Daun Dengan Perbaikan Fitur Citra Menggunakan Metode K-Nearest Neighbor. *Jurnal ULTIMATICS*. **7**(2): 98–104.
- Lombu, A.S., Hidayat, S., & Hidayatullah, A.F. 2022. Pemodelan Klasifikasi Gaji Menggunakan Support Vector Machine. *Journal of Computer System and Informatics (JoSYC)*. **3**(4): 363–70.
- Markoulidakis, I., Rallis, I., Georgulas, I., Kopsiaftis, G., Doulamis, A., & Doulamis, N. 2021. Multi-Class Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem. *Technologies*. **9**(4): 412-9.
- Mitchell, R. 2018. *Web Scraping with Python Collecting Data from the Modern Web*. O'Reilly Media, Inc.
- Naufal, S.A., Adiwijaya, A., & Astuti, W. 2020. Analisis Perbandingan Klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) untuk Deteksi Kanker dengan Data Microarray. *JURIKOM (Jurnal Riset Komputer)*. **7**(1): 162.
- Nurjanah, S., Siregar, A.M., & Kusumaningrum, D.S. 2020. Penerapan Algoritma K-Nearest Neighbor (KNN) Untuk Klasifikasi Pencemaran Udara di Kota Jakarta. *Scientific Student Journal for Information, Technology and Science*. **1**(2): 71–6.
- Prasetyo, E. 2012. *Data Mining Konsep dan Aplikasi Menggunakan Matlab*. Andi, Yogyakarta.
- Purwaningsih, E., & Nurelasari, E. 2021. Penerapan K-Nearest Neighbor Untuk Klasifikasi Tingkat Kelulusan Pada Siswa. *Syntax: Jurnal Informatika*. **10**(01): 46–56.
- Retnoningsih, E., & Pramudita, R. 2020. Mengenal Machine Learning Dengan

- Teknik Supervised Dan Unsupervised Learning Menggunakan Python. *Bina Insani Ict Journal*. 7(2): 156.
- Suguna, N., & Thanushkodi, K. 2010. An Improved K-Nearest Neighbor Classification Using Genetic Algorithm. *International Journal of Computer Science Issues*. 7(4): 18–21.
- Sunarsa, S., & Handayani, R.I. 2016. Sistem Pendukung Keputusan Pemilihan Laptop Untuk Karyawan Pada PT.INDOTEKNO dengan Menggunakan Metode Analytical Hierarchy Process. *Jurnal Ilmu Pengetahuan Dan Teknologi Komputer*. 2(1): 5–10.
- Thandung, D., Lintong, F., & Supit, W. 2013. Tingkat Radiasi Elektromagnetik Beberapa Laptop dan Pengaruhnya Terhadap Keluhan Kesehatan. *Jurnal E-Biomedik (EBM)*. 1(2).
- Turban, E., Aronson, J.E., & Liang, T.P. 2005. *Decision Support System and Intellegent System (Sistem Pendukung Keputusan dan Sistem Cerdas)*. Andi, Yogyakarta
- Wafiyah, F., Hidayat, N., & Perdana, R.S. 2017. Implementasi Algoritma Modified K-Nearest Neighbor (MKNN) untuk Klasifikasi Penyakit Demam. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*. 1(10): 1210–9.
- Wijaya, R., Suciati, N., & Khotimah, W.N. 2017. Implementasi Nearest Neighbor pada Data Kategorik dengan Pembobotan Atribut Menggunakan Weighted Simple Matching Coefficient. *Jurnal Teknik ITS*. 6(2).
- Zhang, Y., Kambhampati, C., Davis, D.N., Goode, K., & Cleland, J.G.F. 2012. A Comparative Study of Missing Value Imputation with Multiclass Classification for Clinical Heart Failure Data. *9th International Conference OnFuzzy Systems and Knowledge Discovery (FSKD 2012)*.
- Zhao, B. 2017. Web Scraping. *Encyclopedia of Big Data*.