PENERAPAN IMPUTASI *K-NEAREST NEIGHBOR* DALAM MENGAMATI KARAKTERISTIK AIR MINUM YANG DIKONSUMSI DI PEDESAAN DAN PERKOTAAN PROVINSI BENGKULU MENGGUNAKAN CHAID

Oleh
AULIA WANDA
NIM. 20337004

PROGRAM STUDI SARJANA STATISTIKA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI PADANG
2024

PENERAPAN IMPUTASI *K-NEAREST NEIGHBOR* DALAM MENGAMATI KARAKTERISTIK AIR MINUM YANG DIKONSUMSI DI PEDESAAN DAN PERKOTAAN PROVINSI BENGKULU MENGGUNAKAN CHAID

SKRIPSI

Diajukan sebagai salah satu persyaratan guna memperoleh gelar
Sarjana Statistika

Oleh
AULIA WANDA
NIM. 20337004

PROGRAM STUDI SARJANA STATISTIKA
DEPARTEMEN STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS NEGERI PADANG

2024

PERSETUJUAN SKRIPSI

PENERAPAN IMPUTASI *K-NEAREST NEIGHBOR* DALAM MENGAMATI KARAKTERISTIK AIR MINUM YANG DIKONSUMSI DI PEDESAAN DAN PERKOTAAN PROVINSI BENGKULU MENGGUNAKAN CHAID

Nama	: Aulia Wanda
NIM	: 20337004
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

Padang, 15 Juli 2024

Mengetahui:
Kepala Departemen Statistika

Dr. Yenni Kurniawati, S.Si., M.Si.
NIP. 198402232010122005

Disetujui Oleh:
Pembimbing

Dr. Yenni Kurniawati, S.Si., M.Si.
NIP. 198402232010122005

PENGESAHAN LULUS UJIAN SKRIPSI

Nama	: Aulia Wanda
NIM	: 20337004
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

PENERAPAN IMPUTASI *K-NEAREST NEIGHBOR* DALAM MENGAMATI KARAKTERISTIK AIR MINUM YANG DIKONSUMSI DI PEDESAAN DAN PERKOTAAN PROVINSI BENGKULU MENGGUNAKAN CHAID

Dinyatakan lulus setelah dipertahankan di depan Tim Penguji Skripsi

Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Negeri Padang

Padang, 15 Juli 2024

Tim Penguji

	Nama	Tanda Tangan
Ketua	: Dr. Yenni Kurniawati S.Si., M.Si	
Anggota	: Dr. Dony Permana, S.Si., M.Si.	
Anggota	: Zilrahmi, S.Pd., M.Si.	

SURAT PERNYATAAN TIDAK PLAGIAT

Saya yang bertanda tangan di bawah ini:

Nama	: Aulia Wanda
NIM	: 20337004
Program Studi	: S1 Statistika
Departemen	: Statistika
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

Dengan ini menyatakan bahwa skripsi saya dengan judul "Penerapan Imputasi *K-Nearest Neighbor* dalam Mengamati Karakteristik Air Minum yang Dikonsumsi di Pedesaan dan Perkotaan Provinsi Bengkulu Menggunakan CHAID" adalah benar merupakan hasil karya saya dan bukan merupakan plagiat dari karya orang lain atau pengutipan dengan cara-cara yang tidak sesuai dan etika yang berlaku dalam tradisi keilmuan. Apabila suatu saat terbukti saya melakukan plagiat maka saya bersedia diproses dan menerima sanksi akademis maupun hukum sesuai dengan hukum dan ketentuan yang berlaku, baik di institusi UNP maupun masyarakat dan negara.

Demikian pernyataan ini saya buat dengan penuh kesadaran dan rasa tanggungjawab sebagai anggota masyarakat ilmiah.

Padang, 15 Juli 2024

Diketahui Oleh:
Kepala Departemen Statistika

Dr. Yenni Kurniawati, S.Si., M.Si.
NIP. 198402232010122005

Saya yang menyatakan

Aulia Wanda
NIM. 20337004

Penerapan Imputasi *K-Nearest Neighbor* dalam Mengamati Karakteristik Air Minum yang Dikonsumsi di Pedesaan dan Perkotaan Provinsi Bengkulu Menggunakan CHAID

Aulia Wanda

ABSTRAK

Air minum merupakan kebutuhan pokok bagi masyarakat selain kebutuhan pangan, sandang, dan papan. Kelayakan air minum perlu diperhatikan supaya aman secara fisik, mikrobiologis, kimia, dan radioaktif. Persentase air minum layak di Indonesia mencapai 91.72% pada tahun 2023. Berdasarkan rata-rata persentase sumber air minum layak antar provinsi dari tahun 2015-2023, Provinsi Bengkulu menempati urutan terendah (56.14%). Akses air minum layak di Provinsi Bengkulu juga berbeda secara signifikan antara pedesaan dengan perkotaan, sehingga penting untuk mengetahui bagaimana karakteristik air minum di kedua wilayah tersebut.

Penelitian ini adalah penelitian terapan menggunakan data air minum Provinsi Bengkulu dari Survei Demografi dan Kesehatan Indonesia tahun 2017. Pada data terdapat masalah *missing data*, sehingga menggunakan imputasi *K-Nearest Neighbor* untuk mengatasinya. Metode untuk analisis adalah metode CHAID.

Tujuan dari penelitian ini adalah untuk mengklasifikasikan karakteristik air minum yang dikonsumsi di pedesaan dan perkotaan Provinsi Bengkulu. Hasil dari metode CHAID menunjukkan bahwa karakteristik air minum di pedesaan adalah air telah direbus dan tidak dibiarkan beberapa waktu dalam wadah/penyimpanan dengan sumber air layak dari tempat lain. Di perkotaan, air minum tidak direbus, tidak disaring dengan penyaring air, tidak dibiarkan beberapa waktu dalam wadah atau penyimpanan, dan air berasal dari rumah sendiri/tempat lain.

Kata Kunci: Air Minum, CHAID, Imputasi *K-Nearest Neighbor*, *Missing Data*.

Application of K-Nearest Neighbor Imputation in Observing the Characteristics of Drinking Water Consumed in Rural and Urban Areas of Bengkulu Province Using CHAID

Aulia Wanda

ABSTRACT

Drinking water is a basic need for the community in addition to food, clothing and shelter. The feasibility of drinking water needs to be considered so that it is physically, microbiologically, chemically, and radioactively safe. The percentage of safe drinking water in Indonesia reached 91.72% in 2023. Based on the average percentage of safe drinking water sources between provinces from 2015-2023, Bengkulu Province had the lowest rank (56.14%). The access to safe drinking water in Bengkulu is also significantly different between rural and urban areas. Therefore, it is important to classify the characteristics of drinking water in both areas.

This study was an applied research using drinking water data of Bengkulu Province from the 2017 Indonesian Demographic and Health Survey. There was a missing data problem, so K-Nearest Neighbor imputation was used to solve it. The method for analysis was the CHAID method.

The objective of this study was to classify the characteristics of drinking water consumed in rural and urban areas of Bengkulu Province. The results of the CHAID method was that the characteristics of drinking water in rural areas were water that had been boiled and not left for some time in a container/storage with a source of safe water from elsewhere. In urban areas, drinking water was not boiled, not filtered with a water filter, not left for some time in a container or storage, and water comes from one's own home/other places.

Keywords: CHAID, Drinking Water, Imputasi K-Nearest Neighbor, Missing data.

KATA PENGANTAR

Puji dan syukur atas limpahan rahmat, nikmat, hidayah, dan inayah yang Allah SWT berikan, sehingga skripsi yang berjudul **"Penerapan Imputasi *K-Nearest Neighbor* dalam Mengamati Karakteristik Air Minum yang Dikonsumsi di Pedesaan dan Perkotaan Provinsi Bengkulu Menggunakan CHAID"** dapat diselesaikan dengan baik. Selawat beserta salam tidak lupa diaturkan kepada Nabi Muhammad SAW.

Penulisan skripsi ini merupakan salah satu syarat yang harus dipenuhi untuk memperoleh gelar Sarjana Statistika pada Program Studi Sarjana Statistika, Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Padang. Skripsi ini dalam penyusunannya tidak terlepas dari berbagai macam kendala, tetapi berkat bantuan, bimbingan, dan dukungan dari berbagai pihak, serta ridho dari Allah SWT sehingga kendala-kendala yang dihadapi dapat diatasi. Oleh sebab itu, ucapan terima kasih disampaikan kepada pihak-pihak yang telah berkontribusi dalam penyusunan skripsi ini.

1. Ibu Dr. Yenni Kurniawati, S.Si., M.Si., selaku dosen pembimbing akademik, pembimbing skripsi dan Kepala Departemen Statistika yang telah meluangkan waktu untuk membimbing dan memberikan arahan selama penyusunan skripsi.
2. Bapak Dr. Dony Permana, S.Si., M.Si., selaku dosen pembahas skripsi yang telah memberikan arahan dan masukan selama penyusunan skripsi.
3. Ibu Zilrahmi, S.Pd., M.Si., selaku dosen pembahas skripsi yang telah memberikan arahan dan masukan selama penyusunan skripsi.

4. Bapak dan Ibu Dosen beserta Staf Departemen Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri Padang.
5. Teristimewa kepada kedua orang tua tercinta, Bapak Amrizal dan Ibu Ermizawati yang senantiasa memberikan doa, semangat, nasihat, materi, serta kasih sayang dan cinta yang tiada batas.
6. Saudara perempuan tercinta, Ai Monica yang selalu siap membantu dan menjadi pendengar terbaik meski terpisah oleh jarak yang jauh. Berbagai macam dukungan, semangat, kasih sayang dan doa selalu tercurah tiada henti.
7. Seluruh anggota Gajah *Girls* yang sudah menemani perjalanan suka dan duka selama di perkuliahan.
8. Serta semua pihak yang telah membantu yang tidak dapat disebutkan satu persatu.

Semoga segala bantuan, bimbingan, dan dukungan yang telah diberikan menjadi amal kebaikan dan mendapat balasan dari Allah SWT.

Pada skripsi ini masih terdapat kekurangan dan jauh dari kesempurnaan. Oleh karena itu, kritik dan saran yang membangun sangat dibutuhkan untuk kesempurnaan skripsi ini. Semoga skripsi ini dapat memberikan manfaat untuk semua pihak yang membutuhkan. *Aamiin*.

Padang,15 Juli 2024

Aulia Wanda

DAFTAR ISI

DAFTAR TABEL

Tabel **Halaman**

DAFTAR GAMBAR

Gambar **Halaman**

DAFTAR LAMPIRAN

Lampiran **Halaman**

BAB I
PENDAHULUAN

A. Latar Belakang Masalah

Air minum merupakan salah satu kebutuhan pokok bagi masyarakat selain kebutuhan pangan, sandang, dan papan. Air minum yang layak untuk dikonsumsi adalah air yang sudah memenuhi syarat dan aman menurut kesehatan (Bambang dkk. 2022). Adapun syarat yang aman menurut kesehatan yaitu aman dari segi fisik, mikrobiologis, kimia, dan radioaktif (Kementerian Kesehatan Republik Indonesia, 2018: 249). Agar memenuhi syarat aman dari segi kesehatan, maka perlu diberi perlakuan terhadap air sebelum diminum. Umumnya, perlakuan yang dilakukan oleh masyarakat yaitu dengan cara direbus, disaring menggunakan kain, diberi bahan kimia, disinfeksi dengan matahari, filtrasi keramik, serta diberi penjernih dan disinfektan (Herlambang, 2010).

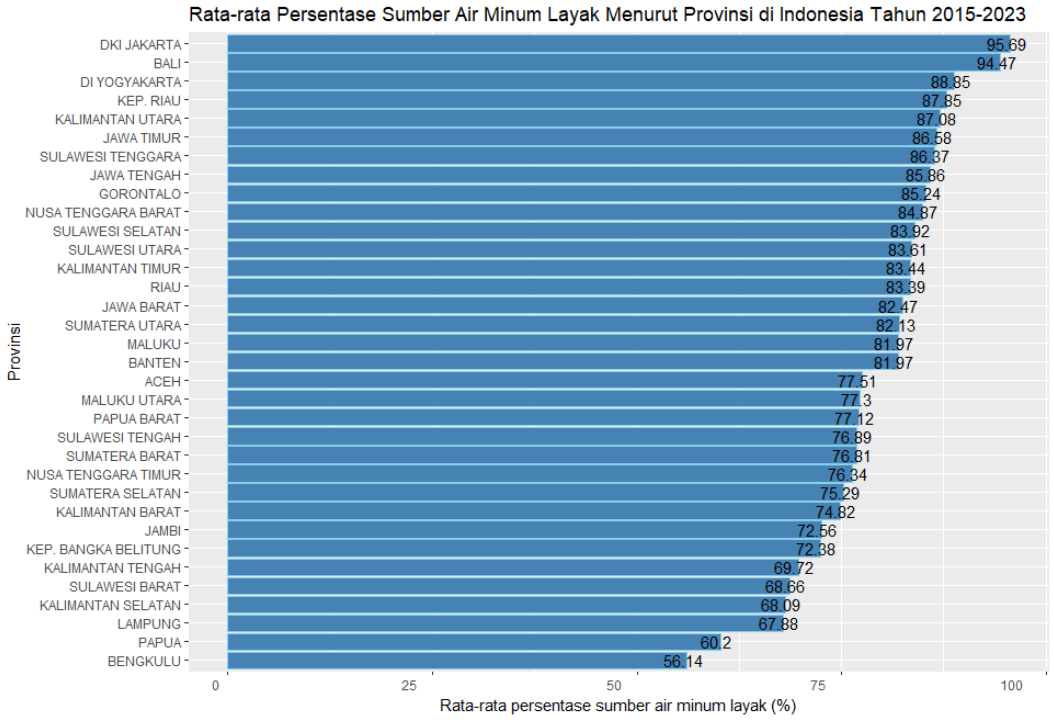
Selain itu, air minum yang layak perlu dilihat darimana sumber airnya berasal. Sumber air minum yang layak tersebut berasal dari sumber air yang bersih. Berdasarkan penelitian yang dilakukan oleh Lestari dan Indriani (2022), sumber air minum yang termasuk ke dalam kategori layak yaitu air pipa tingkat rumah tangga, air pipa umum, air pipa dari tetangga, sumur gali yang terlindungi, sumur bor/pompa, mata air yang terlindungi, air kemasan, dan air isi ulang. Sementara itu, untuk sumber air minum yang termasuk ke dalam kategori tidak layak adalah mata air tidak terlindungi, sumur gali yang tidak terlindungi di tingkat rumah tangga serta umum, sungai, bendungan, danau, air hujan, dan air yang berasal dari truk.

Manfaat yang diperoleh dari mengonsumsi air minum yang layak adalah bisa mencegah tubuh dari penularan berbagai macam penyakit. Penyakit yang bisa menular melalui air yaitu penyakit diare, disentri, tipes, cacingan, kolera, penyakit kulit, dan keracunan (Kementerian Kesehatan Republik Indonesia, 2020). Oleh karena itu, dengan mengonsumsi air minum yang layak menjadi salah satu cara yang dapat dilakukan dalam menjaga kesehatan.

Kenyataan yang terjadi pada masa sekarang untuk akses air minum layak masih belum mencapai target yang diharapkan. Merujuk dari pernyataan *World Health Organization* (WHO) dan *United Nations Children's Fund* (UNICEF) (2021), pada tahun 2020 terdapat sekitar dua miliar masyarakat di dunia belum memperoleh layanan air untuk minum yang terkelola secara aman. Tidak hanya itu, termasuk di dalamnya ada 1,2 miliar orang dengan layanan dasar, 282 juta orang dengan layanan yang terbatas, dan 367 juta orang mengonsumsi sumber air minum tidak layak, serta 122 juta orang minum air permukaan langsung. Kenyataan ini masih jauh dari target yang diharapkan oleh badan Persatuan Bangsa-Bangsa dalam agenda Pembangunan Berkelanjutan tahun 2030 atau dikenal juga dengan *Sustainable Development Goals*/SDGs. Dimana tujuan tentang air terdapat pada *goals* ke-6 dari 17 *goals* yang ada. Isi dari tujuan tersebut adalah seluruh masyarakat di dunia dapat mengakses air bersih dan sanitasi yang layak yang terjamin ketersediaannya (Pertiwi, 2023).

Jika dilihat di Indonesia, berdasarkan data Survei Sosial Ekonomi atau SUSENAS pada tahun 2023, persentase rumah tangga dengan sumber air minum layak adalah sebesar 91.72% (Badan Pusat Statistik, 2023). Capaian akses air

minum layak ini berbeda cukup jauh dengan tahun 2015-2018, dimana persentase rumah tangga yang memiliki sumber air minum layak masih berada di bawah 80%. Jika dilihat dari rata-rata persentase sumber air minum layak menurut provinsi dari tahun 2015-2023 yang terdapat pada Gambar 1, Provinsi Bengkulu menempati posisi paling terendah (56.14%).



Gambar 1. Rata-rata Persentase Sumber Air Minum Layak Menurut Provinsi di Indonesia Tahun 2015-2023

Persentase dari akses air minum layak di Provinsi Bengkulu selain menempati posisi terendah juga memiliki perbedaan yang cukup jauh untuk wilayah perkotaan dengan pedesaannya (Badan Pusat Statistik, 2023). Rumah tangga yang tinggal di wilayah perkotaan memiliki kecenderungan yang lebih tinggi untuk menggunakan air minum layak daripada yang tinggal di pedesaan (Nurzanah dkk., 2020; Putri dan Yuniasih, 2022; Rahim dan Muchlisoh, 2020). Hal ini didukung oleh penelitian yang dilakukan oleh Lestari dan Indriani (2022), dimana di wilayah perkotaan akses

air minum yang layak dan aman jauh lebih mudah diperoleh. Dalam penelitian tersebut juga mengungkapkan bahwa antara sumber air minum dengan jenis wilayah tempat tinggal (perkotaan dan pedesaan) memiliki hubungan yang signifikan. Selain itu, pengolahan sumber air minum sebelum diminum juga memiliki hubungan yang signifikan dengan jenis wilayah tempat tinggal. Maka, pada penelitian ini perlu digali lebih lanjut agar bisa diketahui apa saja karakteristik dari air minum yang dikonsumsi oleh rumah tangga untuk wilayah pedesaan dan perkotaan di Provinsi Bengkulu.

Berdasarkan pemaparan di atas, untuk bisa mengetahui bagaimana karakteristik air minum yang dikonsumsi di pedesaan dan perkotaan Provinsi Bengkulu, digunakan data Survei Demografi dan Kesehatan Indonesia (SDKI) tahun 2017. SDKI merupakan survei yang dilakukan oleh beberapa badan, yaitu Badan Pusat Statistik (BPS), Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN), dan Kementerian Kesehatan (Kemenkes). Tujuannya yaitu menyediakan estimasi yang paling baru dari indikator dasar demografi dan kesehatan di Indonesia. SDKI menyajikan estimasi level nasional dan provinsi dengan cakupan 1.970 blok sensus yang termasuk di dalamnya wilayah perkotaan dan pedesaan. Kemudian, pada pelaksanaannya menggunakan empat jenis kuesioner yaitu kuesioner rumah tangga, wanita usia subur (WUS), pria kawin (PK), dan remaja pria (RP) (Badan Kependudukan dan Keluarga Berencana Nasional dkk., 2018: 1-2). Pada data SDKI memuat keadaan rumah tangga, termasuk tentang air minum yang dikonsumsi sehari-hari.

Data yang berasal dari survei tidak selalu lengkap atau memiliki *missing data* (data hilang), begitu juga dengan data SDKI. Beberapa penelitian terdahulu yang menggunakan data SDKI mengungkapkan bahwa dalam data tersebut memiliki *missing data* (Marfuqoh dan Martha, 2020; Situmeang dkk., 2017; Tambing dkk., 2023; Yogo dan Wahyuni, 2019). *Missing data* adalah permasalahan yang terjadi pada suatu gugus data yang menyebabkan beberapa bagian dari data hilang atau tidak lengkap. Penyebabnya yaitu terdapat responden yang menolak untuk bekerja sama secara langsung, menolak menjawab beberapa pertanyaan dengan alasan tertentu, dan kehilangan data karena kesalahan dalam penginputan hasil wawancara atau kesalahan teknis lainnya (Longford, 2005: 13).

Missing data menurut Little dan Rubin (2020: 13-14) memiliki tiga mekanisme, yaitu *Missing Completely at Random* (MCAR), *Missing at Random* (MAR), dan *Missing not at Random* (MNAR). MCAR adalah *missing data* yang terjadi pada gugus data tidak berkaitan dengan nilai data, baik itu dengan yang hilang maupun yang teramati. Kemudian, MAR merupakan *missing data* yang terjadi pada gugus data berkaitan dengan komponen yang diamati atau lainnya. Sementara itu, MNAR adalah *missing data* yang terjadi pada gugus data yang kaitannya yaitu dengan komponen-komponen yang hilang tersebut. Masalah yang terjadi karena *missing data* ini bisa menjadi keterbatasan dari penelitian yang dilakukan. Ada dua akibat yang ditimbulkan, yaitu hilangnya efisiensi data dan bias (Carpenter dkk., 2023: 10). Maka, sangat perlu dilakukan penanganan *missing data* yang sesuai dengan mekanismenya agar tidak kehilangan efisiensi data dan bias.

Proses untuk menggali berbagai macam pengetahuan dan informasi dari suatu gugus data biasanya dikenal dengan nama *data mining*. *Data mining* merupakan kegiatan yang dilakukan terhadap suatu gugus data yang sebelumnya tidak punya arti penting, dengan mencari secara berulang serta intensif untuk memperoleh pengetahuan dari gugus data tersebut (Pramana dkk., 2023: 2). Menurut Manurung dan Hasugian (2019), secara fungsionalitas, tugas *data mining* dikelompokkan menjadi 6 kelompok, yaitu klasifikasi, klasterisasi, regresi, deteksi anomali, pembelajaran aturan asosisasi, dan perangkuman. Pada penelitian ini akan melakukan proses *data mining* guna memperoleh informasi-informasi yang terdapat dalam data SDKI, dimulai dari tahapan eksplorasi untuk mengidentifikasi *missing data* dan penanganannya, hingga menganalisis gugus data tersebut.

Sebagaimana yang diungkapkan oleh para peneliti terdahulu bahwa di dalam data SDKI terdapat masalah *missing data*, maka perlu untuk diberikan penanganan. Umumnya, penanganan *missing data* yang dilakukan oleh peneliti-peneliti, yaitu *listwise deletion*, *pairwise deletion*, dan *imputation* (Raudhatunnisa dan Wilantika, 2021). *Listwise deletion* merupakan cara penanganan *missing data* dimana semua objek yang mengandung data hilang akan dihapus. Kemudian, *pairwise deletion* adalah penanganan yang dilakukan dengan menghapus data yang hilang tanpa menghapus unit yang lainnya dalam baris yang terdapat hilang tersebut. Metode *imputation* atau dikenal dengan imputasi merupakan penanganan *missing data* yang dilakukan dengan mengganti nilai yang hilang menggunakan nilai yang memungkinkan dari gugus data tersebut. Metode imputasi dinilai lebih baik dibanding penanganan *listwise deletion* dan *pairwise deletion*. Hal ini dikarenakan

pada kedua metode tersebut dapat membuat hasil perhitungan menjadi tidak valid karena menghapus amatan atau nilai yang hilang yang menyebabkan informasi penting lainnya menjadi berkurang (Handayany dkk., 2023).

Metode imputasi terdiri dari dua jenis, yaitu berbasis statistik dan *machine learning*. Menurut Raudhatunnisa dan Wilantika (2021), imputasi berbasis statistik memakai aturan statistik dalam menjalankan proses imputasi, sedangkan imputasi berbasis *machine learning* memanfaatkan pelatihan pada data sehingga bisa memprediksi nilai yang hendak diimputasi. Dalam penelitian Handayany dkk. (2023) menyebutkan contoh dari imputasi berbasis statistik adalah imputasi *Mean*, *Regression*, *Hot-Deck*, dan *Multiple Imputation*. Kemudian, untuk imputasi berbasis *machine learning* adalah imputasi *Multilayer Perceptron*, *Self Organization Maps*, C4.5*,* CN2, dan *K-Nearest Neighbor*.

Metode imputasi yang umumnya dipakai adalah imputasi berbasis *machine learning,* yaitu imputasi *K-Nearest Neighbor*. Jika dibandingkan dengan metode imputasi yang lain, imputasi *K-Nearest Neighbor* dinilai lebih baik dari segi kualitas akurasinya (Fadillah dan Muchlisoh, 2020; Handayany dkk., 2023; Yusuf dkk., 2023). Imputasi *K-Nearest Neighbor* dilakukan dengan melihat amatan terdekat dan kemiripan nilai yang hilang pada gugus data (Sallaby dan Azlan, 2021). Imputasi ini mampu digunakan untuk penanganan *missing data* dengan tipe data numerik atau kategorik serta tidak memerlukan asumsi (Handayany dkk., 2023). Selain itu, imputasi *K-Nearest Neighbor* juga dapat digunakan untuk menangani *missing data* dengan mekanisme MCAR, MAR, dan juga MNAR (Yusuf dkk., 2023). Kelemahannya adalah membutuhkan waktu yang lama ketika mencari amatan

untuk menangani amatan yang hilang pada data berukuran cukup besar, karena akan mencarinya secara keseluruhan dari data *training* atau gugus data (Sudrajat dan Cholid, 2023).

Setelah dilakukan proses eksplorasi dan teridentifikasi terdapat *missing* data, lanjut dengan penanganan untuk *missing data*, sehingga dihasilkan gugus data yang siap untuk dianalisis. Adapun metode analisis yang bisa dipakai untuk menentukan karakteristik air minum yang dikonsumsi di pedesaan dan perkotaan Provinsi Bengkulu adalah metode klasifikasi. Klasifikasi adalah pengelompokan yang dilakukan secara sistematis ke dalam kelas tertentu dengan melihat ciri-ciri yang sama/mirip (Helena dkk., 2019). Metode klasifikasi terbagi menjadi dua, yaitu parametrik dan nonparametrik. Klasifikasi parametrik umumnya disertai dengan asumsi tertentu yang telah ditentukan sebelum melakukan analisis (Hartono dkk., 2020). Berbeda dengan klasifikasi nonparametrik yang tidak bergantung pada asumsi tertentu dan lebih mudah dalam proses analisis data, serta tetap memiliki nilai akurasi yang tinggi dengan memperhatikan metode klasifikasi yang digunakan adalah tepat (Anugrah dkk., 2022). Selain itu, klasifikasi nonparametrik dinilai lebih siap dalam menghadapi berbagai macam kondisi data. Hal ini disebabkan oleh keterbatasan yang dimiliki oleh klasifikasi parametrik karena ketika asumsinya tidak sesuai dengan kondisi data, maka perlu dilakukan penanganan untuk memenuhi asumsi tersebut (Hartono dkk., 2020).

Klasifikasi nonparametrik terbagi lagi menjadi beberapa metode, yaitu *Decision Tree*, *Naive Bayes*, *Neural Network*, *Random Forest*, dan lain sebagainya (Ardiansyah dkk., 2018). Salah satu metode yang banyak digunakan dari beberapa

metode klasifikasi tersebut adalah *Decision Tree*. *Decision Tree* atau dikenal juga dengan pohon keputusan merupakan metode klasifikasi yang mampu mendeteksi serta menghitung efek nonlinear pada variabel respons dan interaksi diantara variabel prediktor (Juwita dkk., 2021). *Decision Tree* terdiri dari beberapa metode, yaitu *Classification and Regression Tree* (CART), *Chi-square Automatic Interaction Detection* (CHAID), dan *Quick Unbiased Efficient Statistical Tree* (QUEST), dan lain sebagainya (Fajriati dan Syafriandi, 2022).

Salah satu metode *Decision Tree* yang umum dipakai adalah metode CHAID. Metode CHAID merupakan bagian dari metode *Automatic Interaction Detection* (AID) yang menggunakan uji statistik *chi-square* (De Ville, 2006: 35). Metode CHAID dapat digunakan jika variabel prediktornya berskala nominal atau ordinal (Amalita dkk., 2019). Hasil dari analisisnya adalah berupa pohon klasifikasi/diagram pohon yang mudah untuk diinterpretasikan (Zulaiha dkk., 2023). Metode CHAID bekerja dengan mempelajari hubungan antara variabel respons dengan prediktor yang kemudian mengklasifikasikan sampel berdasarkan hubungan tersebut (Rakhmawati dkk., 2023).

Metode CHAID memiliki keunggulan dalam mengeksplorasi data yang berjumlah besar dengan semua tipe variabelnya adalah kategorik (Anugrah dkk., 2022). Kelebihan lain yang dimiliki oleh CHAID adalah tepat sasaran, mampu mendefinisikan ke dalam kelas regu yang sesuai, dan iteratif (Fajriati dan Syafriandi, 2022). Selain itu, dari beberapa penelitian terdahulu untuk tingkat akurasinya ketika dibandingkan dengan beberapa metode klasifikasi lainnya menunjukkan bahwa CHAID memiliki akurasi yang baik.

Beberapa penelitian terdahulu yang menyatakan bahwa metode CHAID memiliki akurasi yang baik adalah seperti penelitian yang dilakukan oleh Permana dkk. (2021) yang berjudul Komparasi Performa Algoritma ID3, C4.5, CHAID dalam Profiling Tersangka Kasus Narkoba di Jawa Barat, dalam hasilnya menyatakan bahwa akurasi CHAID paling tinggi dari ketiga metode tersebut sehingga digunakan algoritma CHAID untuk mendapatkan model terbaiknya. Selanjutnya penelitian Faisal dkk. (2017) yang berjudul Perbandingan Kinerja Metode Klasifikasi *Chi-square Automatic Interaction Detection* (CHAID) dengan Metode Klasifikasi Algoritma C4.5 pada Studi Kasus: Penderita Diabetes Melitus Tipe 2 Di Samarinda Tahun 2015. Pada penelitian tersebut diperoleh bahwa hasil akurasi dari metode CHAID lebih tinggi daripada metode C4.5. Kemudian, penelitian yang dilakukan oleh Sa'diah dkk. (2021) yang berjudul Klasifikasi Pemberian Kredit Sepeda Motor Menggunakan Metode Regresi Logistik Biner dan *Chi-square Automatic Interaction Detection* (CHAID) dengan GUI R (Studi Kasus: Kredit Sepeda Motor di PT X). Hasil penelitiannya menunjukkan bahwa ketepatan klasifikasi metode CHAID lebih baik daripada Regresi Logistik Biner dalam mengklasifikasikan kredit macet di perusahaan X.

Pada penelitian ini menggunakan metode CHAID untuk mengetahui karakteristik dari air minum yang dikonsumsi oleh rumah tangga di pedesaan dan perkotaan Provinsi Bengkulu. Namun, sebelum dianalisis, datanya terlebih dahulu diberi penanganan untuk *missing data* yaitu menggunakan metode imputasi *K-Nearest Neighbor*. Maka, penelitian ini diberi judul **"Penerapan Imputasi *K-Nearest Neighbor* dalam Mengamati Karakteristik Air Minum yang**

Dikonsumsi di Pedesaan dan Perkotaan Provinsi Bengkulu Menggunakan CHAID".

B. Batasan Masalah

Berdasarkan masalah yang telah dipaparkan pada latar belakang, perlunya batasan masalah untuk membuat penelitian ini menjadi terarah. Adapun batasan masalah dalam penelitian ini adalah data yang digunakan yaitu data SDKI tentang air minum yang dikonsumsi oleh rumah tangga di pedesaan dan perkotaan Provinsi Bengkulu tahun 2017.

C. Rumusan Masalah

Berdasarkan latar belakang masalah di atas, dapat dirumuskan masalah untuk penelitian ini adalah sebagai berikut:

1. Bagaimana asosiasi antara variabel respons dengan prediktor melalui uji *chi-square*?
2. Bagaimana karakteristik air minum yang dikonsumsi oleh rumah tangga di pedesaan dan perkotaan Provinsi Bengkulu menggunakan metode CHAID dengan penanganan *missing data* imputasi *K-Nearest Neighbor*?

D. Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Mengetahui asosiasi antara variabel respons dengan prediktor melalui uji *chi-square*.
2. Mengetahui karakteristik air minum yang dikonsumsi oleh rumah tangga di pedesaan dan perkotaan Provinsi Bengkulu menggunakan metode CHAID dengan penanganan *missing data* imputasi *K-Nearest Neighbor*.

E. Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat kepada:

1. Peneliti, menambah wawasan dan pengetahuan tentang proses *data mining*, penanganan *missing data* dengan metode imputasi *K-Nearest Neighbor* serta analisis menggunakan metode CHAID.
2. Pemerintah, sebagai informasi tentang karakteristik air minum yang dikonsumsi oleh rumah tangga di wilayah pedesaan dan perkotaan Provinsi Bengkulu. Diharapkan dengan mengetahui hal tersebut pemerintah terutama pemerintah Provinsi Bengkulu dapat mengambil kebijakan yang tepat untuk memberikan edukasi kepada masyarakat serta penanganan yang lebih lanjut dalam menyukseskan harapan tercapainya akses air minum yang layak secara merata baik di perkotaan maupun pedesaan.
3. Peneliti selanjutnya, sebagai referensi yang dapat dipakai untuk melanjutkan penelitian ini dengan cakupan yang lebih luas lagi.