MODEL K-NEAREST NEIGHBOR YANG EFISIEN UNTUK PREDIKSI PENYAKIT JANTUNG

Oleh
YOLANDA DJAFAR
T3114090

SKRIPSI

PROGRAM SARJANA
TEKNIK INFORMATIKA
UNIVERSITAS ICHSAN GORONTALO
GORONTALO
2018

JUDUL SKRIPSI

MODEL K-NEAREST NEIGHBOR YANG EFISIEN UNTUK PREDIKSI PENYAKIT JANTUNG

Oleh
YOLANDA DJAFAR
T3114090

SKRIPSI

PROGRAM SARJANA
TEKNIK INFORMATIKA
UNIVERSITAS ICHSAN GORONTALO
GORONTALO
2018

PERSETUJUAN SKRIPSI

MODEL K-NEAREST NEIGHBOR YANG EFISIEN UNTUK PREDIKSI PENYAKIT JANTUNG

Oleh
YOLANDA DJAFAR
T3114090

SKRIPSI

Untuk memenuhi salah satu syarat ujian
guna memperoleh gelar Sarjana
Program Studi Teknik Informatika,
ini telah disetujui oleh Tim Pembimbing

Gorontalo, November 2019

Pembimbing I	Pembimbing II
Budy Santoso, S.Kom., M.Eng	**Yulianty Lasena, S.Kom., M.Kom**

PENGESAHAN SKRIPSI

MODEL K-NEAREST NEIGHBOR YANG EFISIEN UNTUK PREDIKSI PENYAKIT JANTUNG

Oleh
YOLANDA DJAFAR
T3114090

Diperiksa oleh Panitia Ujian Strata Satu (S1)
Universitas Ichsan Gorontalo

1. Ketua Penguji
   Nama Penguji I ……..

2. Anggota
   Nama Penguji II ……..

3. Anggota
   Nama Penguji III ……..

4. Anggota
   Budy Santoso, S.Kom., M.Eng ……..

5. Anggota
   Yulianty Lasena, S.Kom., M.Kom ……..

PERNYATAAN SKRIPSI

Dengan ini saya menyatakan bahwa:

1. Karya tulis (skripsi) saya ini adalah asli dan belum pernah diajukan untuk mendapatkan gelar akademik (Sarjana) baik di Universitas Ichsan Gorontalo maupun di perguruan tinggi lainnya.
2. Karya tulis (skripsi) saya ini adalah murni gagasan, rumusan, dan penelitian saya sendiri, tanpa bantuan pihak lain, kecuali arahan dari Tim Pembimbing.
3. Dalam karya tulis (skripsi) saya ini tidak terdapat karya atau pendapat yang telah dipublikasikan orang lain, kecuali secara tertulis dicantumkan sebagai acuan/sitasi dalam naskah dan dicantumkan pula dalam daftar pustaka.
4. Peryataan ini saya buat dengan sesungguhnya dan apabila dikemudian hari terdapat penyimpangan dan ketidakbenaran dalam peryataan ini, maka saya bersedia menerima sanksi akademik berupa pencabutan gelar yang telah diperoleh karena karya tulis ini, serta sanksi lainnya sesuai dengan norma-norma yang berlaku di Univeristas Ichsan Gorontalo.

Gorontalo, November 2019

Yang membuat pernyataan,

Yolanda Djafar

ABSTRAK

Kematian yang disebabkan penyakit jantung masih sangat tinggi, sehingga perlu peningkatan upaya-upaya pencegahannya, misalnya dengan meningkatkan capaian model prediksinya. Penerapan metode-metode *machine learning* pada *dataset* publik (*Cleveland*, *Hungary*, *Switzerland*, *VA Long Beach*, & *Statlog*) yang umumnya digunakan oleh para peneliti untuk prediksi penyakit jantung, termasuk pengembangan alat bantunya, masih belum menangani *missing value*, *noisy data*, *unbalanced class*, dan bahkan *data validation* secara efisien. Oleh karena itu, pendekatan imputasi *mean/mode* diusulkan untuk menangani *missing value replacement*, *Min-Max Normalization* untuk menangani *smoothing noisy data*, *K-Fold Cross Validation* untuk menangani *data validation*, dan pendekatan *ensemble* menggunakan metode *Weighted Vote* (WV) yang dapat menyatukan kinerja tiap-tiap metode *machine learning* untuk mengambil keputusan klasifikasi sekaligus untuk mereduksi *unbalanced class*. Hasil penelitian ini menunjukkan bahwa metode yang diusulkan tersebut memberikan akurasi sebesar 85,21%, sehingga mampu meningkatkan kinerja akurasi metode-metode *machine learning*, selisih 7,14% dengan *Artificial Neural Network*, 2,77% dengan *Support Vector Machine*, 0,34% dengan C4.5, 2,94% dengan *Naïve Bayes*, dan 3,95% dengan *k-Nearest Neighbor*.

KATA PENGANTAR

Alhamdulillah, penulis panjatkan kehadirat Allah Subhanahu wa Ta'ala atas berkat rahmat dan hidayah-Nya, sehingga penulis dapat menyelesaikan usulan penelitian ini untuk memenuhi salah satu syarat penyusunan Skripsi Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Ichsan Gorontalo.Salam dan taslim kepada junjungan kita, Nabi Muhammad Shallallahu 'Alaihi wa Sallam atas perjuangan beliau yang telah mengantar kita dari alam kebodohan ke alam yang penuh ilmu pengetahuan. Usulan penelitian ini penulis beri judul: **"Model K-Nearest Neighbor yang Efisien untuk Prediksi Penyakit Jantung."**

Penulis menyadari sepenuhnya bahwa usulan penelitian ini tidak mungkin terwujud tanpa bantuan dan dorongan dari berbagai pihak, baik bantuan moril maupun materil. Untuk itu, dengan segala keikhlasan dan kerendahan hati, penulis mengucapkan terima kasih dan penghargaan yang setinggi-tingginya kepada:

1. Ibu Dra. Hj. Juriko Abdussamad, M.Si. selaku Ketua Yayasan Pengembangan Ilmu Pengetahuan dan Teknologi (YPIPT) Ichsan Gorontalo;
2. Bapak Dr. Abdul Gaffar La Tjokke, M.Si., selaku Rektor Universitas Ichsan Gorontalo;
3. IbuZohrahayaty, M.Kom., selaku Dekan Fakultas Ilmu Komputer Universitas Ichsan Gorontalo;
4. Ibu Asmaul Husna N., M.Kom., selaku Wakil Dekan I Bidang Akademik;
5. Ibu Irma Surya Kumala, M.Kom., selaku Wakil Dekan II Bidang Kepegawaian, Administrasi Umum, dan Keuangan;
6. Bapak Yasin Aril Mustofa, M.Kom., selaku Wakil Dekan III Bidang Kemahasiswaan.
7. Bapak Irvan Abraham Salihi, M.Kom., selaku Ketua Program Studi Teknik Informatika Fakultas Ilmu Komputer Universitas Ichsan Gorontalo;
8. Bapak Budy Santoso, S.Kom., M.Eng., selaku Dosen Pembimbing Utama yang telah membimbing penulis dalam penyusunan usulan penelitian ini;
9. Ibu Yulianty Lasena, S.Kom., M.Kom., selaku Dosen Pembimbing Pedamping yang telah membimbing penulis dalam penyusunan usulan penelitian ini;

10. Bapak dan Ibu Dosen Universitas Ichsan Gorontalo yang telah mendidik dan mengajarkan berbagai disiplin ilmu kepada penulis;
11. Kedua Orang Tua dan keluarga atas segala kasih sayang, jerih payah dan doa restunya dalam membesarkan, mendidik penulis, serta memberikan bantuan dan dukungan moril yang sangat besar kepada penulis;
12. Rekan-rekan seperjuangan yang telah banyak memberikan bantuan dan dukungan moril yang sangat besar kepada penulis;
13. Semua pihak yang ikut membantu dalam penyelesaian usulan penelitian ini yang tak sempat penulis sebutkan satu-persatu.

Semoga Allah, SWT melimpahkan balasan atas jasa-jasa mereka dankepada penulis. Selanjutnya, penulis menyadari sepenuhnya bahwa apa yang telah dicapai ini masih jauh dari kesempurnaan dan masih banyak terdapat kekurangan. Oleh karena itu, penulis sangat mengharapkan adanya kritik dan saran yang konstruktif.

Akhirnya penulis berharap semoga hasil yang telah dicapai ini dapat mendukung program pemerintah dalam mencerdaskan kehidupan bangsa serta bermanfaat bagi pengembangan ilmu pengetahuan dan teknologi khususnya di bidang ilmu komputer maupun bermanfaat bagi masyarakat.

Gorontalo, November 2019

Yolanda Djafar

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR LAMPIRAN

BAB I
PENDAHULUAN

1.1 Latar Belakang

Data *World Health Organization* (WHO) pada tahun 2012 menunjukkan 17.5 juta jiwa (31%) penduduk di dunia meninggal akibat penyakit jantung dan jumlah tersebut akan terus meningkat [1], [2]. Sementara data *The Institute for Health Metrics and Evaluation* (IHME) pada tahun 2016 menunjukkan 17.7 juta jiwa (32.26%) penduduk di dunia meninggal karena penyakit jantung [3]. Survei dari *Sample Regristration System* pada tahun 2014 di Indonesia menunjukkan Penyakit Jantung Koroner berada pada posisi tertinggi setelah Stroke sebagai penyebab kematian, yaitu sebesar 12.9% [1].

Dengan demikian, perlu peningkatan upaya-upaya pencegahannya, seperti deteksi dini penyakit jantung yang dapat dilakukan melalui diagnosa oleh dokter. Namun masyarakat umumnya malas melakukan cek kesehatan secara berkala, minimnya pengetahuan mereka mengenai penyakit jantung, dan kendala biaya dapat mempersulit deteksi dini penyakit jantung [4]. Alternatif alat bantu untuk prediksi penyakit jantung dapat menjadi salah satu solusi dari kesulitan tersebut. Oleh karena itu, peningkatan capaian model prediksi penyakit jantung perlu terus ditingkatkan sehingga dapat digunakan untuk pengembangan alat bantunya.

Pendekatan *Machine Learning* (ML) memiliki potensi besar untuk mengatasi masalah-masalah dalam domain biomedis komputasi [5], seperti prediksi penyakit jantung. *Dataset* publik yang umumnya digunakan oleh para peneliti dalam menemukan model ML yang paling akurat untuk prediksi penyakit jantung tersedia di *UCI Machine Learning Repository* (*Cleveland*, *Hungary*, *Switzerland*, *VA Long Beach*, & *Statlog*) [6]. Kelima *dataset* tersebut memiliki karakteristik variabel yang sama (secara rinci ditunjukkan pada Tabel 3.2), sehingga dapat disatukan.

Kecuali *Statlog dataset*, keempat *dataset* lainnya memiliki masalah *Missing Value* (MV) dan *Unbalanced Class* (UC) yang memang sering terjadi pada data klinis harus dikendalikan dengan sangat hati-hati, karena sangat berefek pada hasil prediksi [7], [8]. Sementara adanya *outlier* atau anomali pada data dapat

menyebabkan *Noisy Data* (ND) atau data yang tidak konsisten [7], [9], berdampak buruk terhadap kinerja model [10]. Begitupun dengan *Data Validation* (DV), tekniknya mesti tepat, partisi data latih dan data uji harus menjamin bahwa semua sampel digunakan untuk pelatihan dan pengujian model, setiap label *class* mestinya terwakili secara merata pada pelatihan model [9]. Secara rinci, masalah MV dan UC pada lima *dataset* prediksi penyakit jantung ditunjukkan pada Tabel 3.1.

Namun penerapan metode-metode ML dan varian-variannya pada *dataset* penyakit jantung masih kurang memperhatikan masalah MV, ND, UC, dan bahkan DV. Sehingga walaupun akurasi yang tinggi telah dicapai, namun belum efisien secara keseluruhan. Begitupun dari sisi pengembangan aplikasinya, tiga aplikasi berbasis *web* yang telah dikembangkan menggunakan metode ML (*Naïve Bayes*) yang justru akurasinya lebih rendah dan bahkan sangat tidak memperhatikan masalah MV, ND, UC, dan DV pula. Secara rinci, kinerja akurasi dari penerapan metode-metode ML yang diusulkan untuk prediksi penyakit jantung maupun pengembangan aplikasinya ditunjukkan pada Tabel 2.1. Sedangkan studi-studi survei pada topik prediksi penyakit jantung ditunjukkan pada Tabel 2.2.

Membuang MV bisa menghilangkan informasi yang mungkin penting, mengakibatkan *bias* [11]. Pendekatan imputasi merupakan strategi yang efisien untuk menangani MV [9], [11]. Selanjutnya, cara klasik dan umum digunakan untuk mereduksi ND, yaitu pendekatan normalisasi [10]. Sementara itu, untuk menangani masalah UC, dapat dilakukan dengan cara menyatukan kelima dataset tersebut.

Dengan meningkatkan efisiensi dan kinerja model prediksi penyakit jantung melalui penanganan *Missing Value Replacement* (MVR) menggunakan pendekatan inputasi *average*, *Smoothing Noisy Data* (SND) menggunakan pendekatan *Min-Max Normalization*, *Unbalanced Class Reduction* (UCR) dengan pendekatan manual, dan *Data Validation* menggunakan metode *K-Fold Cross Validation*, maka diharapkan dapat meningkatkan kinerja model prediksi penyakit jantung. Sementara metode klasifikasi yang digunakan untuk model prediksi penyakit jantung adalah *K-Nearest Neighbor* (K-NN), karena atribut-atribut dataset penyakit jantung bertipe *integer*, *real*, *ordinal*, dan *binominal*.

Dengan demikian, berdasarkan berbagai uranian tersebut, maka peneliti tertarik mengusulkan penelitian dengan formulasi judul: **"Model *K-Nearest Neighbor* yang Efisien untuk Prediksi Penyakit Jantung."** Diharapkan metode yang diusulkan ini dapat digunakan untuk pengembangan sistem cerdas deteksi dini penyakit jantung, sehingga berdampak dalam mereduksi tingkat kematian yang disebabkan penyakit jantung dan dapat mempermudah masyarakat.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah diuraikan, maka dapat diidentifikasi masalah umum dan masalah komputasi penelitian ini sebagai berikut:

1. Kematian yang disebabkan penyakit jantung sangat tinggi dan sulit dideteksi karena merupakan penyakit yang termasuk dalam kategori *silent killer*. Deteksi dini (prediksi) penyakit jantung dapat dilakukan melalui diagnosa yang tepat oleh dokter. Namun masyarakat umumnya jarang melakukan cek kesehatan secara berkala, minimnya pengetahuan masyarakat mengenai penyakit jantung, dan kendala biaya, maka diperlukan suatu alat bantu untuk mempermudah deteksi/prediksi dini penyakit jantung bagi masyarakat. Oleh karena itu, model prediksi penyakit jantung perlu terus ditingkatkan untuk pengembangan alat bantunya.
2. Pendekatan *Machine Learning* (ML) memiliki potensi besar untuk mengatasi masalah-masalah dalam domain biomedis komputasi, seperti prediksi penyakit jantung. Namun dataset publik yang umumnya digunakan oleh para peneliti dalam menemukan model ML yang paling akurat untuk prediksi penyakit jantung masih belum menangani masalah UC, MV, ND, dan DV secara efisien. Dengan cara manual, UCR dapat diatasi. Membuang MV dapat menyebabkan bias, imputasi sederhana *mean* (*average*) pada atribut numerik dan *mode* pada atribut ordinal/binominal dapat digunakan untuk MVR. Untuk meningkatkan konsistensi data, pendekatan *Min-Max Normalization* dapat digunakan untuk SND. Pada dataset yang berukuran cukup besar, pendekatan yang efisien digunakan untuk melakukan DV adalah *K-Fold Cross Validation*. Selanjutnya,

K-NN dapat digunakan sebagai metode klasifikasi karena atribut-atribut dataset penyakit jantung yang bertipe *integer*, *real, ordinal*, dan *binominal*.

1.3 Rumusan Masalah

Berdasarkan identifikasi masalah, maka dapat dirumuskan masalah penelitian ini, yaitu: Bagaimana peningkatan kinerja Model *K-Nearest Neighbor* untuk Prediksi Penyakit Jantung apabila efisiensi pra pengolahan data ditingkatkan melalui *Unbalanced Class Reduction* secara manual, *Missing Value Replacement* menggunakan pendekatan *average/mode*, *Smoothing Noisy Data* menggunakan *Min-Max Normalization*, dan *Data Validation* menggunakan *K-Fold Cross Validation*?

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah, maka tujuan penelitian ini adalah: Meningkatkan kinerja Model *K-Nearest Neighbor* untuk Prediksi Penyakit Jantung melalui pra pengolahan data yang lebih efisien yang terdiri dari *Unbalanced Class Reduction* secara manual, *Missing Value Replacement* menggunakan pendekatan *average/mode*, *Smoothing Noisy Data* menggunakan *Min-Max Normalization*, dan *Data Validation* menggunakan *K-Fold Cross Validation*.

1.5 Manfaat Penelitian

Dampak yang dapat terjadi apabila tujuan penelitian ini tercapai, yaitu:

1. Manfaat teoritis, memberikan konstribusi ke ilmu pengetahuan, khususnya pada bidang ilmu komputer, berupa Model *K-Nearest Neighbor* yang efisien untuk Prediksi Penyakit Jantung.
2. Manfaat praktis, dengan peningkatan capaian prediksi penyakit jantung melaui model K-Nearest Neighbor yang lebih efisien, maka diharapkan dapat digunakan untuk pengembangan sistem cerdas deteksi dini penyakit jantung yang kinerjanya lebih baik sehingga berdampak dalam mereduksi kematian yang disebabkan penyakit jantung dan mempermudah masyarakat.

BAB II
LANDASAN TEORI

2.1 Tinjauan Studi

Aakurasi 100% ditunjukkan *Hidden Naïve Bayes – Inter Quartile Range* pada *dataset Statlog* yang tidak memiliki masalah MV dan UC [12]. *Hidden Naïve Bayes* berfungsi sebagai metode klasifikasi, namun mentransformasi data yang sudah terstruktur menjadi data *image* menggunakan *Inter Quartile Range* justru menyebabkan adanya MV yang selanjutnya mereka ganti dengan nilai *mean/mode*. Mentransformasi data yang sudah terstruktur menjadi tidak terstruktur itu justru malah bisa meningkatkan ND. Penelitian tersebut mengejar akurasi tapi mengorbankan efisiensi secara keseluruhan, karena justru menyebabkan *bias* pada data. Sementara penelitian-penelitian terkait lainnya masih belum menangani masalah UC, MV, dan DV secara efisien. Secara rinci, penerapan metode-metode *Machine Learning* (ML) untuk prediksi penyakit jantung ditunjukkan pada Tabel 2.1, sementara studi-studi serveinya ditunjukkan pada Tabel 2.2 berikut ini.

Tabel 2.1 Literatur Studi (Technical Papers)

Year, Ref.	Dataset	Methods	MV	UC	DV	Accuracy
2008, [13]	Cleveland	NB	?	☑	⊗	95%/web
2010, [14]	Cleveland	DT–GA	?	☑	☑	99,2%
2011, [15]	Cleveland	DT–GA	?	☑	☑	99,2%
2012, [16]	Cleveland & Statlog	ANN	☑	?	?	99,25%
2012, [17]	?	NB	?	?	?	web
2013, [4]	Cleveland	NB	?	☑	⊗	95%/web
2015, [18]	Statlog	ANN	☑	☑	?	85%
2016, [12]	Statlog	HNB–IQR	⊗	☑	☑	100%
2016, [19]	Cleveland	J48	☑	?	☑	56,76%
2016, [20]	Cleveland	k-NN–Apriori	⊗	?	☑	99,19%
2016, [21]	Semua kecuali Satlog	Ensemble–ML	⊗	?	☑	92%
2016, [22]	Cleveland	Rule methods	☑	?	☑	86,7%
2017, [23]	Hungary	DT	?	?	☑	67,7%
2018, [24]	Cleveland	RF	☑	?	☑	83,15%
2018, [25]	Cleveland	ICO	⊗	?	?	84,17%
2018, [26]	Cleveland	Ensemble–ML	⊗	?	?	90%

Keterangan:
☑ Tepat/baik atau tidak perlu ditangani
? Tidak diketahui atau tidak teridentifikasi
⊗ Kurang/tidak tepat, tidak ditangani, diabaikan, atau dibuang

Tabel 2.2 Litaratur Studi (Review Paper)

Year, Ref.	Authors	Proposed Method
2012, [27]	Shouman, Turner and Stocker	DT – GA = 99,2%
2012, [28]	Vijayarani	SVM
2016, [29]	Banu and Swamy	DT = 99,62 acc
2017, [30]	Gnaneswar and Jebarani	ANN
2017, [31]	Sowmiya and Sumitra	ML – Apriori

2.2 Tinjauan Pustaka

2.2.1 Data Mining

Data merupakan entitas yang tidak memiliki arti, meskipun kemungkinan memiliki nilai di dalamnya. Tentunya data perlu disimpan, namun yang lebih penting dari itu adalah bagaimana menggali atau menemukan pengetahuan dari data yang disimpan. Data merupakan kumpulan/rekaman/catatan dari fakta, transaksi, atau objek tentang suatu kejadian yang tidak membawa arti. Dapat pula merupakan suatu catatan terstruktur dari suatu transaksi. Sedangkan informasi merupakan data yang telah diolah sehingga memiliki nilai. Dengan demikian, data merupakan materi penting dalam membentuk informasi. Sementara itu, pengetahuan merupakan gabungan dari informasi yang bertujuan untuk memberikan suatu informasi yang baru. Pengetahuan dapat berupa solusi pemecahan suatu masalah atau petunjuk suatu pekerjaan, dimana nilainya bisa ditingkatkan, dipelajari, dan diajarkan kepada orang lain.

Sejak tahun 1990, *Data Mining* sudah mulai dikenal, di saat pemanfaatan atau pengolahan data menjadi sesuatu yang penting dalam berbagai pekerjaan dan bidang, baik itu di bidang akademis, bisnis, medis, dsb [32]. Dengan demikian usia *Data Mining* masih terbilang remaja, sehingga posisinya dalam disiplin ilmu masih diperdebatkan. Dalam jurnal ilmiah, *Data Mining* dikenal dengan nama *Knowladge Discovery in Databases* (KDD). Beberapa sebutan lain *Data Mining*, antara lain: *Knowledge Extraction*, Analisis Data, *Pattern Recognition*, serta *Business Intelligence*.

Teknologi *database* saat ini telah memungkinkan penyimpanan data dalam jumlah yang besar dan terakumulasi. Besarnya jumlah data yang tersimpan inilah yang melatar belakangi adanya *Data Mining*. Tentunya data perlu disimpan, namun yang lebih penting dari itu adalah bagaimana menggali atau menemukan pengetahuan dari data yang disimpan.

Mengapa *Data Mining* penting? Data yang sedemikian besar tentunya memiliki informasi yang tersembunyi di dalamnya, namun kemampuan manusia terbatas dalam menganalisis data atau menemukan informasi yang tersembunyi atau menemukan pola di dalam data yang berjumlah besar. Menemukan informasi yang tersembunyi atau pola di dalam data yang berjumlah besar ini dapat pula diistilahkan dengan mengekstrak pengetahuan yang tentunya sangat berguna untuk mendukung pengambilan kebijakan atau keputusan. Selain itu, kemampuan komputasi yang semakin canggih dan terjangkau, serta persaingan bisnis yang semakin kompetitif merupakan faktor-faktor lainnya mengapa *Data Mining* perlu dilakukan.

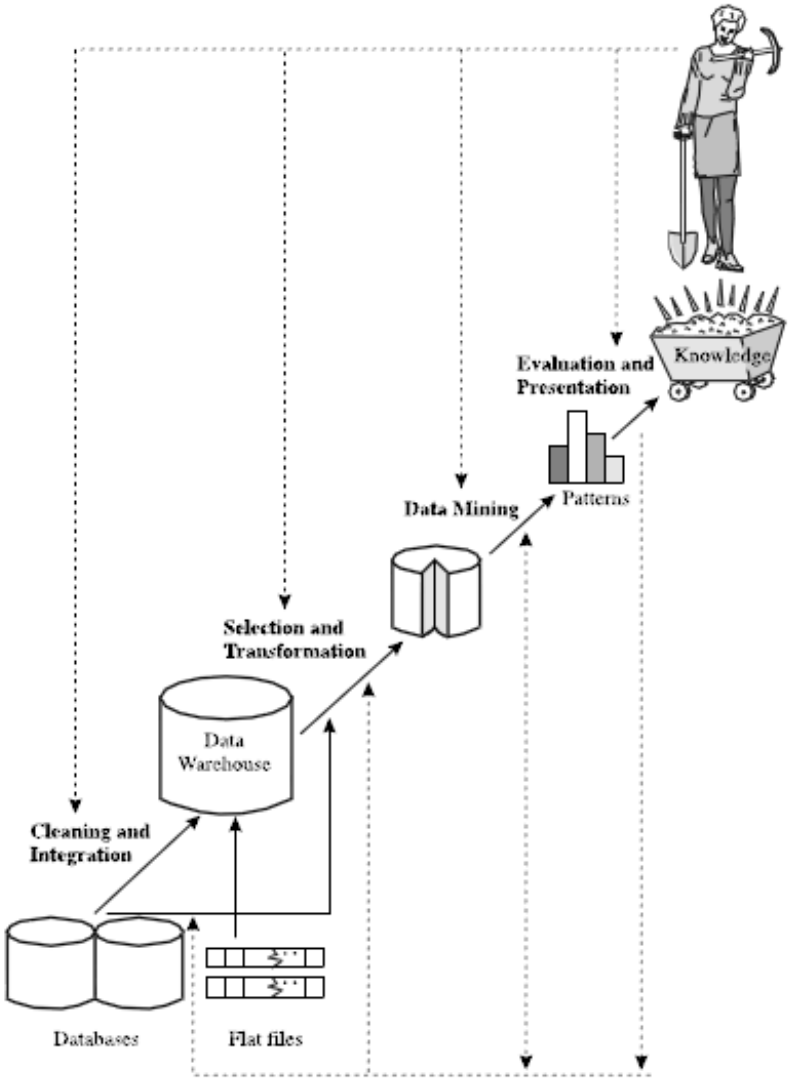
Meskipun *Data Mining* dapat diartikan juga sebagai penemuan informasi, namun tidak semua penemuan informasi itu merupakan *Data Mining* [33]. Contohnya pencarian informasi masakan di Google, bukan merupakan *Data Mining*, tapi pengelompokkan masakan berdasarkan negara pada hasil pencarian di Google, merupakan *Data Mining*. Definisi *Data Mining* menurut J. Han, M. Kamber, & J. Pei (2012) adalah: "*The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner* [34]."

Berikut ini adalah tahapan proses dalam *Data Mining* [34]:

1. Pembersihan data. Dilakukan untuk menghilangkan noise pada data dan data yang tidak konsisten.
2. Integrasi data. Dilakukan untuk menggabungkan beberapa sumber data.
3. Seleksi data. Melakukan seleksi terhadap data, di mana data yang relevan dengan metode analisis *Data Mining* yang dilakukan diambil dari database.
4. Transformasi data. Melakukan transformasi terhadap data, di mana data ditransformasikan atau dikonsolidasikan ke dalam bentuk yang sesuai dengan

tugas analisis atau agar dapat dianaliss oleh metode *Data Mining* yang digunakan. Hal ini dapat dilakukan dengan operasi agregasi atau dengan akumulasi.

5. Metode *Data Mining*. Penerapan metode *Data Mining* dalam mengekstrak data untuk mengenal pola dari data. Hasil dari proses ini dapat berupa suatu pengetahuan atau model atau pola.
6. Evaluasi model/pola. Mengukur atau mengevaluasi atau mengidentifikasi pola-pola yang mewakili pengetahuan yang diperoleh.
7. Presentasi pengetahuan. Memberikan pengetahuan yang diperoleh kepada pengguna.



Gambar 2.1 Tahapan Proses Data Mining

Data dalam terminologi statistik adalah kumpulan objek dengan atribut-atribut tertentu, di mana objek tersebut adalah individu berupa data dan setiap data memiliki sejumlah atribut. Semakin banyak atribut, maka semakin besar pula dimensi dari data tersebut. Kumpulan data-data membentuk dataset. Data dapat pula diistilahkan dengan vektor, dapat pula diistilahkan dengan *record*.

Terdapat 3 kategori dataset, yaitu [33]:

- *Record*: data matriks, data transaksi, data dokumen, atau data yang sudah terstruktur.
- *Graph*: *Word Wide Web* (WWW), struktural molekul, *image*, *voice*, dsb.
- Dataset lainnya: data spasial, data temporal, data sekuensial, data urutan *genetic*, dsb.

Jenis dataset ada 2, yaitu [33]:

- *Dataset private*. Dataset yang diambil dari organisasi yang dijadikan obyek masalah/kasus, misalnya dari suatu bank, rumah sakit, industri, pabrik, perusahaan, dsb. Pengetahuan yang diperoleh dari proses Data Mining dengan menggunakan dataset private tentunya akan bersifat privasi pula, tidak *comparable* dan *verifiable*.
- *Dataset public*. Dataset dapat diambil dari repositori publik yang disepakati oleh para peneliti Data Mining, misalnya: UCI Repository, ACM KDD Cup, dataset yang diambil dari berbagai perusahaan yang sejenis, dsb. Pengetahuan yang diperoleh dari proses Data Mining dengan menggunakan dataset public akan bersifat: *comparable*, *repeatable*, dan *verifiable*.

Atribut atau variabel terbagi menjadi 2, yaitu [33]:

- Atribut input atau variabel input atau faktor atau parameter atau indikator.
- Atribut output atau variabel output atau target atau *class* atau *label* (kategori dari class).

Walaupun secara praktis tipe data untuk atribut pada *Data Mining* hanya menggunakan 2 tipe, yaitu nominal (diskrit) dan numerik (kontinyu atau ordinal), namun lebih lengkapnya tipe data untuk atribut terbagi menjadi 4 seperti yang ditunjukkan pada Tabel 2.3.

Tabel 2.3 Tipe Atribut [33]

TIPE		PENJELASAN	CONTOH
Kategorikal (Kualitatif)	Nominal	Nilai atribut bertipe nominal memberikan nilai berupa nama, dengan nama inilah sebuah atribut membedakan dirinya pada vektor yang satu dengan yang lain.	Kode pos, nomor KTP, jenis kelamin.
	Ordinal	Nilai atribut bertipe ordinal mempunyai nilai berupa nama yang mempunyai arti informasi yang terurut.	Predikat (cum laude, sangat puas, puas), suhu (dingin, normal, panas).
Numerik (Kuantitatif)	Interval	Nilai atribut di mana perbedaan di antara dua nilai mempunyai makna yang berarti.	Tanggal, suhu (dalam Celcius atau Fahrenheit).
	Rasio	Nilai atribut di mana perbedaan di antara dua nilai dan rasio dua nilai mempunyai makna yang berarti.	Umur, panjang, tinggi, rata-rata.

Secara umum, tugas *Data Mining* terbagi 2, yaitu: (1) Deskriptif: Mengenal karakter dari data; dan (2) Prediktif: Melakukan prediksi berdasarkan pola atau pengetahuan yang diperoleh dari pembelajaran terhadap data [34]. Sedangkan secara rinci, tugas *Data Mining* terbagi menjadi 5, yaitu [34], [35]:

1. *Clustering*. Pendekatan ini berupaya menemukan hubungan antara atribut input sebagai target/output, sehingga tidak memiliki varibael *output/target/class*. Suatu *cluster* adalah kumpulan dari catatan/data yang mirip satu sama lainnya, dan berbeda dengan catatan dalam kelompok lainnya. Tugas *Clustering* tidak mencoba untuk mengklasifikasikan atau memprediksi nilai dari atribut output namun mencari segmen seluruh data yang ditetapkan menjadi sub kelompok yang relatif *homogen/cluster*, di mana kesamaan catatan/data dalam *cluster* dimaksimalkan dan kesamaan *cluster* di luar catatan/data diminimalkan. Clustering dapat dijadikan sebagai langkah awal dalam proses *Data Mining*, di mana *cluster* yang dihasilkan digunakan sebagai masukan lebih lanjut (dijadikan sebagai variabel output) ke teknik/tugas lainnya yang berbeda, misalnya ke tahap/tugas *classification* atau *estimation*. Beberapa algoritma *clustering* yang umum digunakan: *K-Means, K-Medoids, Self-Organizing Map, Fuzzy C-Means,* dsb.
2. *Classification*. Pendekatan ini berusaha mememukan hubungan antara atribut input dengan atribut output. Atribut output inilah yang disebut sebagai guru

atau yang melakukan pengawasan terhadap proses yang dilakukan. Dengan demikian, pendekatan ini memiliki atribut *output/target/class*. Hasil *classification* merupakan suatu pengetahuan/pola yang digunakan untuk memprediksi nilai dari atribut *output/terget/class* yang bertipe kategorikal/nominal dari nilai-nilai atribut inputan yang belum diketaui nilai outputnya/targetnya. Kata lainnya, memprediksi kasus baru berdasarkan kasus yang ada atau pengalaman yang telah terjadi. Beberapa algoritma *classification* yang umum digunakan: *ID3, C4.5, CART, Naive Bayes, K-Nearest Neighbor, Linear Discriminant Analysis, Artificial Neural Network, Support Vector Machine,* dsb.

3. *Estimation/Regression*. Pendekatan ini mirip dengan *classification*, perbedaannya adalah pada atribut *output/target/class*. Atribut output pada *classification* bertipe kategorikal/nominal, sedangkan pada *estimation* bertipe numerikal. Beberapa algoritma estimation yang umum digunakan: *Artificial Neural Network, Support Vector Machine, Linear Regression*, dsb.
4. *Association*. Pendekatan ini berusaha menemukan atribut input yang muncul bersamaan dalam suatu transaksi. Dalam dunia bisnis, sering disebut dengan affinity analysis atau *market basket analysis*. Pendekatan ini mencari aturan yang menghitung hubungan diantara dua atau lebih item data. Berangkat dari pola "*If antecedent, then consequent*," bersamaan dengan pengukuran *support* (coverage) dan *confidence* (accuration) yang terasosiasi dalam aturan. Dengan demikian, pendekatan ini hanya memiliki 1 atribut input dan tidak memiliki atribut *output/target/class*. Beberapa algoritma *association* yang umum digunakan: *A-Priori, FP-Growth, dan GRI*.

2.2.2 Machine Learning

Machine Learning merupakan salah satu teknik dalam bidang *Artificial Intelligence* (AI), selain teknik *optimization, searching,* dan *reasoning*. *Machine Learning* mengambil keputusan dengan cara malakukan pembelajaran terhadap objek atau pengalaman yang telah terjadi, sehingga bisa saja tidak memerlukan pengetahuan atau aturan atau kepakaran seperti pada teknik reasoning. Dengan

begitu, Teknik ini bahkan terkadang mampu memberikan pengetahuan baru. Metode yang sering digunakan dalam *Data Mining* adalah metode-metode *Machine Learning*, seperti: *Artificial Neural Network, Support Vector Machine, K-Nearest Neighbor, Naïve Bayes, ID3, C 4.5, K-Means,* dsb. Walaupun demikian, pekerjaan *Data Mining* biasanya tidak hanya melibatkan metode-metode *Machine Learning*. Agar dapat memberikan model *Data Mining* yang kinerjanya lebih baik, biasanya dikombinasikan dengan metode-metode AI lainnya, misalnya kombinasi metode learning *Artificial Neural Network* dengan metode reasoning *Fuzzy Logic* menghasilkan metode *Adaptive Neuro Fuzzy Inferences System* (ANFIS), dsb.

Metode *Machine Learning* dapat dikategorikan menjadi 2, yaitu [33]:

1. *Unsupervised learning*. Metode ini melakukan pembelajaran tanpa pengawasan, atau biasa diistilahkan dengan pembelajaran tanpa guru. Metode ini berusaha mememukan hubungan data antara atribut input (pola/pengetahuan/aturan) untuk menentukan atribut target/class, sehingga tidak memiliki atribut *target/class*. Tugas *Data Mining* yang termasuk *unsupervised learning*, yaitu *clustering* dan *association*. Namun pada *association*, bukan berusaha menemukan hubungan data antara atribut input karena hanya memiliki 1 atribut input saja, tetapi berusaha menemukan hubungan antara item data dari 1 atribut input tersebut. Beberapa algoritma *unsupervised learning*, yaitu: *K-Means, K-Medoids, Self-Organizing Map, Fuzzy C-Means, A-Priori, FP Growth*, dsb.
2. *Supervised learning*. Metode ini melakukan pembelajaran dengan pengawasan, atau biasa diistilahkan dengan pembelajaran dengan guru. Metode ini berusaha mememukan hubungan data antara atribut input dengan atribut *output/target/class*. Atribut ouput inilah yang disebut sebagai guru atau yang melakukan pengawasan terhadap pembelajaran yang dilakukan dalam menemukan pola, pengetahuan, atau aturan untuk memprediksi nilai dari atribut output dari nilai-nilai atribut inputan yang belum diketaui nilai outputnya. Kata lainnya, memprediksi kasus baru dengan cara mempelajari pengalaman (kasus yang ada). Tugas *Data Mining* yang termasuk *supervised learning*, yaitu *classification* dan *estimation*. Beberapa algoritma *supervised*

learning, yaitu: *Naïve Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Random Forest, Linear Discriminant Analysis, Artificial Neural Network, Support Vector Machine*, dsb.

2.2.3 Data Pre-Processing

MV yang jumlahnya sedikit dapat diatasi dengan cara membuangnya. Namun jika MV jumlahnya banyak, seperti pada beberapa *dataset* prediksi penyakit jantung, maka membuangnya dapat menghilangkan informasi yang mungkin penting, mengakibatkan *bias* [11]. Pendekatan imputasi merupakan salah satu strategi yang dapat digunakan untuk mengatasi MV [9], [11]. Imputasi secara sederhana dapat dilakukan dengan mengganti MV menjadi nilai *mean* (1) pada variabel numerik dan nilai *mode* pada variabel kategorikal [7], [9], [11], [36].

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

Adanya *outlier* atau anomali pada data dapat menyebabkan ND [7], [9], berdampak buruk terhadap kinerja model [10]. Intinya, *outlier* mestinya dibuang dengan cara mendeteksinya lebih dahulu, misalnya dengan model prediksi. Namun cara lain yang klasik dan dapat diterapkan untuk mereduksi ND yaitu dengan melakukan normalisasi data, salah satunya menggunakan pendekatan *Min-Max Normalization* (2) [10].

$$X_{ij}^{*} = \frac{(X_{ij} - X_{\min j})}{(X_{\max j} - X_{\min j})}\left(\left(New_{\max j} - New_{\min j}\right) + New_{\min j}\right) \tag{2}$$

Strategi DV bukan hanya sekedar untuk memperoleh akurasi yang tinggi. Metode yang dapat menjamin bahwa semua sampel digunakan untuk pelatihan dan pengujian model, begitupun setiap label *class* terwakili secara merata pada pelatihan model, yaitu *K-Fold Cross Validation* dan *Leave One Out* [9]. Ketika ukuran sampel cukup besar, *K-Fold Cross Validation* adalah pilihan terbaik [9], [37].

2.2.4 K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* (K-NN) menghitung jarak atau kemiripan antar data, menggunakan Persamaan (3).

$$y' = \underset{v}{argmax} \sum_{i=1}^{n} x_i y_i \in D_z I(v = y_i) \tag{3}$$

Perhitungan jarak dapat mengunakan metode *Euclidean* yang ditunjukkan pada Persamaan (4), *Manhattan*, atau lainnya. Simpan hasilnya di dalam *D*, dan pilih $D_z \in D$ yang merupakan *K* tetangga terdekat dari *z*. Sementara *v* merupakan jumlah data yang masuk dalam *class* y_i.

$$d(X_{ij}) = \sqrt{\sum_{k=1}^{n} (X_{ik} - X_{jk})^2} \tag{4}$$

Pembobotan untuk menghitung kandidat *class* yang sebaiknya menjadi hasil prediksi diperoleh melalui Persamaan (5).

$$W_i = \frac{1}{d(x', x_i)^2} \tag{5}$$

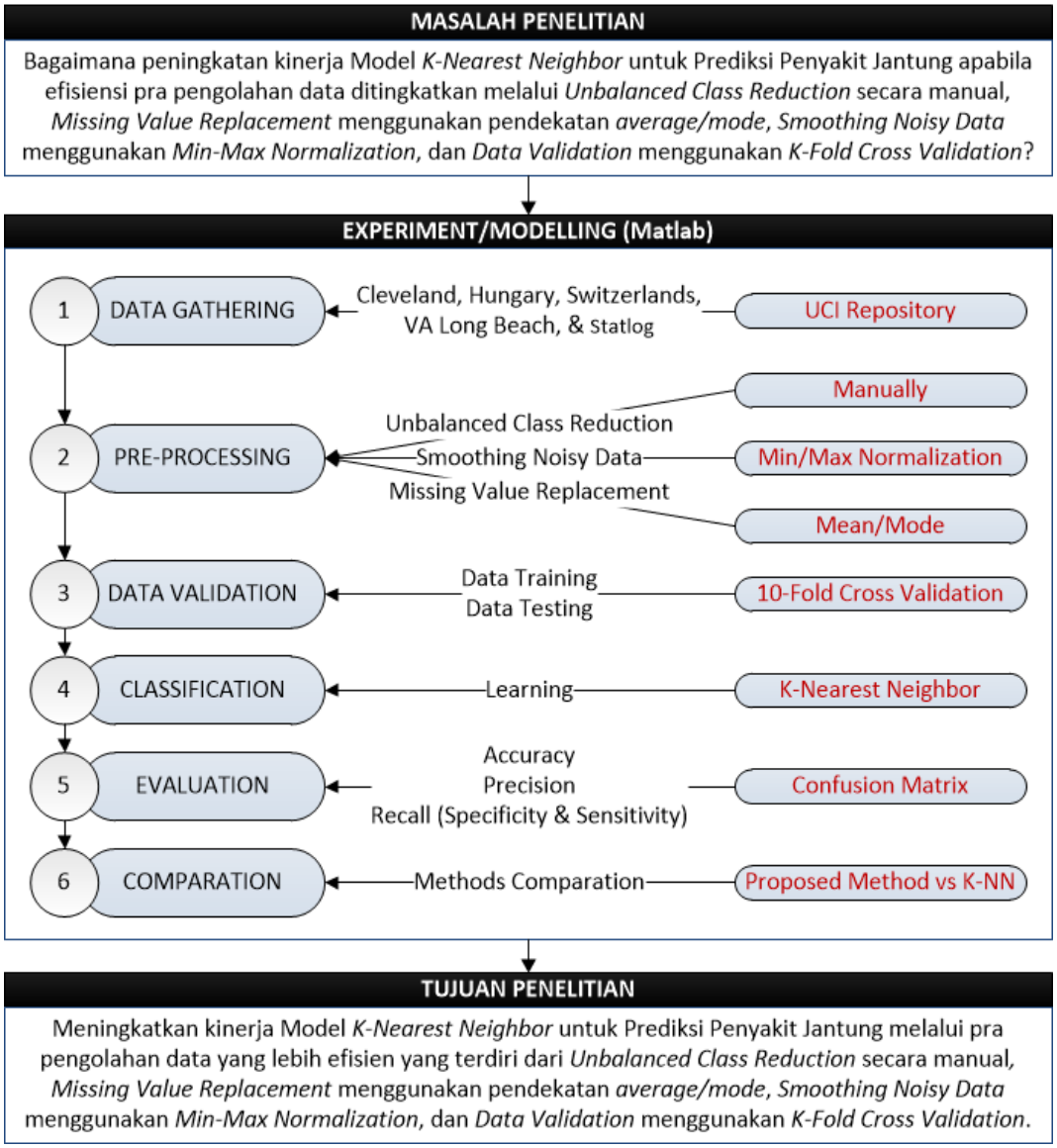
2.2.5 Evaluasi Model

Pengukuran kinerja dari suatu model klasifikasi/prediksi dapat dilakukan menggunakan pendekatan *Confussion Matrix* untuk memperoleh *accuracy*, *precision*, *recall* (*sensitivity* dan *specificity*), dan *F-Measure*.

Tabel 2.4 Confusion Matrix

	Actual Positive	Actual Negative	Precision
Predicted Positive	**True Positive (TP)**	False Positive (FP)	TP / (TP + FP) *
Predicted Negative	False Negative (FN)	**True Negative (TN)**	TN / (TN + FN)
Recall	TP / (TP + FN)	TN / (TN + FP)	
F-Measure	(2 * Precision * Sensitivity) / (Precision + Sensitivity)		
Accuracy	= (TP + TN) / (TP + TN + FN + FP)		

2.3 Kerangka Pikir



Gambar 2.2 Kerangka Pikir

BAB III
METODE PENELITIAN

3.1 Jenis, Subjek, Objek, dan Waktu Penelitian

Dipandang dari tingkat penerapannya, maka penelitian ini merupakan penelitian terapan. Dipandang dari jenis informasi yang diolah, maka penelitian ini merupakan penelitian kuantitatif. Dipandang dari perlakuan terhadap data, maka penelitian ini merupakan penelitian konfirmatori. Penelitian ini menggunakan metode penelitian eksperimen. Dengan demikian jenis penelitian ini adalah penelitian eksperimental.

Subjek penelitian ini adalah pendekatan klasifikasi dalam Data Mining dengan menggunakan metode Machine Learning pada objek prediksi Penyakit Jantung. Penelitian ini dimulai dari bulan September 2018 hingga bulan November 2019.

3.2 Pengumpulan Data

Pengumpulan data primer menggunakan teknik dokumentasi, yaitu dengan mengunduh data publik penyakit jantung pada *UCI Machine Learning Repository*, yaitu dataset: *Cleveland*, *Hungary*, *Switzerlands*, *VA Long Beach* [6], & *Statlog* [38]. Karakteristik kelima dataset tersebut ditunjukkan pada Tabel 3.1. *Class* (variabel output) terdiri dari dua label, yaitu negatif dan positif [6], [38]. Sedangkan variabel inputnya ditunjukkan pada Tabel 3.2.

Tabel 3.1 Karakteristik Dataset Penyakit Jantung

Code	Dataset	Instances	MV	UC
D1	Cleveland	303	6	No
D2	Hungary	294	782	Yes
D2	Switzerland	123	273	Yes
D4	VA Long Beach	200	698	Yes
D5	Statlog	270	0	No
	Total	1,190	1,759	Yes

Tabel 3.2 Variabel Input Dataset Penyakit Jantung [6], [38]

Variable	Type	Description
Umur	Integer	29 – 77 tahun
Jenis Kelamin	Binominal	0 (Wanita); 1 (Pria)
Jenis Sakit Dada	Ordinal	1 (Typical Angina); 2 (Atypical Angina); 3 (Non Anginal Pain); 4 (Asymptomatic)
Tekanan Darah	Integer	80 – 200
Kolestrol	Integer	85 – 603
Kadar Gula	Binominal	0 (<=120mg/dl); 1 (>120mg/dl)
Elektrokardiografi	Ordinal	0 (Normal); 1 (Kelainan ST-T); 2 (Hypertrofi Ventrikel)
Tekanan Jantung	Integer	60 – 202
Angina Induksi	Binominal	0 (No); 1 (Yes)
Oldpeak	Real	0 – 6,2
Slope	Ordinal	1 (Naik); 2 (Datar); 3 (Turun)
Denyut jantung	Ordinal	0 – 3
Thal	Ordinal	3 (Normal); 6 (Cacat Tetap); 7 (Cacat Reversibel)

3.3 Unbalanced Class Reduction

Masalah UC yang terjadi pada semua dataset kecuali *Statlog dataset* diatasi dengan menggabungkan data yang termasuk *class* 1 (ada indikasi positif penyakit jantung), 2 (>1), 3 (>2), dan 4 (positif penyakit jantung) menjadi *class* 2, sementara *class* 0 menjadi *class* 1, sehingga akhirnya antara *class* 1 dan 2 menjadi seimbang. Hal ini juga membuat *Statlog dataset* yang hanya memiliki *class* 1 (negatif penyakit jantung) dan 2 (positif penyakit jantung) dan tidak memiliki masalah UC dapat digabungkan dengan dataset lainnya.

3.4 Missing Value Replacement

Membuang MV dapat menyebabkan *bias*. Oleh karena itu, MVR dilakukan dengan mengganti MV menjadi nilai *mean* (1) pada variabel numerik (*integer* dan

real). Sedangkan pada variabel *ordinal* dan *binominal*, MV diganti dengan nilai *mode*.

3.5 Smothing Noisy Data

Adanya *outlier* atau anomali pada data dapat menyebabkan ND yang berdampak buruk terhadap kinerja model. *Outlier* mestinya dibuang dengan cara mendeteksinya lebih dahulu atau mereduksinya melalui SND. Dalam melakukan SND, pendekatan yang digunakan adalah *Min-Max Normalization* (2).

3.6 Data Validation

Strategi *Data Validation* (DV) menggunakan metode: *10-Fold Cross Validation* (90% data latih dan 10% data uji), di mana data dipartisi menjadi 10 bagian/sampel, setiap bagian digunakan untuk data uji, 10 kali secara bergiliran, sisanya sebagai data latih. Metode *Leave One Out* memang lebih efisien karena menguji setiap sampel, namun mempertimbangkan ukuran sampel yang cukup besar, maka metode *K-Fold Cross Validation* yang juga tidak kalah efisien menjadi pilihan.

3.7 Classification (K-Nearest Neighbor)

Setelah pra pengolahan data, selanjutnya dilakukan klasifikasi dengan melatih algoritma *Machine Learning* (K-NN) untuk memperoleh model K-NN untuk prediksi Penyakit Jantung yang paling akurat. Adapun parameter-parameter K-NN yang diuji coba ditunjukkan pada Tabel 3.3 berikut ini.

Tabel 3.3 Parameter K-NN yang Diuji Coba

Parameter	**Nilai yang Diuji Coba**
K	2 s/d n-1, dimana n adalah jumlah data latih dalam K-Fold.
Distance measure	Euclidean, Cityblock, Correlation, & Cosine.
Vote rule	Nearest.

3.8 Evaluasi Model

Kinerja model K-NN untuk prediksi Penyakit Jantung yang diusulkan ini dievaluasi berdasarkan *accuracy*, *precision*, *sensitivity*, dan *specificity* menggunakan pendekatan *Confusion Matrix*.

3.9 Komparasi

Komparasi dilakukan untuk membandingkan kinerja model K-NN untuk prediksi Penyakit Jantung yang diusulkan ini dengan kinerja model K-NN tanpa melalui pra pengolahan data.

BAB IV
HASIL PENELITIAN

4.1 Hasil Eksperimen

4.1.1 Hasil Pra Pengolahan Data

Setelah pengumpulan data, langkah pertama yang dilakukan, yaitu UCR dengan pendekatan secara manual (*dataset mixing*) seperti ditunjukkan pada Tabel 4.1 berikut ini.

Tabel 4.1 Variabel Output (y)

Dataset	**UC**		**UCR**	
D1	0 (Sehat)	54,13%	0 (Negative)	54,13%
	1 (Resiko Rendah)	18,15%	1 ← 1, 2, 3, 4 (Positive)	45,87%
	2 (Resiko Sedang)	11,88%		
	3 (Resiko Tinggi)	11,55%		
	4 (Resiko Sangat Tinggi)	4,29%		
D2	0 (Sehat)	63,95%	0 (Negative)	63,95%
	1 (Sakit)	36,05%	1 (Positive)	36,05%
D3	0 (Sehat)	6,50%	0 (Negative)	6,50%
	1 (Resiko Rendah)	39,02%	1 ← 1, 2, 3, 4 (Positive)	93,50%
	2 (Resiko Sedang)	26,02%		
	3 (Resiko Tinggi)	24,39%		
	4 (Resiko Sangat Tinggi)	4,07%		
D4	0 (Sehat)	25,50%	0 (Negative)	25,50%
	1 (Resiko Rendah)	28,00%	1 ← 1, 2, 3, 4 (Positive)	74,50%
	2 (Resiko Sedang)	20,50%		
	3 (Resiko Tinggi)	21,00%		
	4 (Resiko Sangat Tinggi)	5,00%		
D5	1 (Negatif)	55,56%	0 (Negative)	55,56%
	2 (Positif)	44,44%	1 (Positive)	44,44%
All			**0 (Negative)**	**41,13%**
			1 (Positive)	**58,87%**

Label output (*class*) tiap-tiap *dataset* berbeda-beda namun memiliki karakteristik yang sama. Dengan demikian, masalah UC yang terjadi pada semua *dataset* kecuali *Statlog dataset* diatasi dengan menggabungkan data yang termasuk *class* 1 (ada indikasi positif penyakit jantung), 2 (>1), 3 (>2), dan 4 (positif penyakit jantung) menjadi *class* 1, sementara *class* 0 tetap *class* 0, sehingga akhirnya antara *class* 0 dan 1 menjadi seimbang. Hal ini juga membuat *Statlog dataset* yang hanya memiliki *class* 1 (negatif penyakit jantung) dan 2 (positif penyakit jantung) yang

tidak memiliki masalah UC dapat digabungkan dengan *dataset* lainnya dengan cara yang sama.

Setelah UCR, selanjutnya dilakukan MVR dengan mengganti MV menjadi nilai *mean* (1) pada variabel numerik (*integer* dan *real*) dan nilai *mode* pada variabel *ordinal*/*binominal*. Berikut ini adalah fungsi MVR yang dibuat menggunakan alat bantu Matlab:

```
function [datasetMV] = aaMVR(data) % MVR
   [n,d] = size(data); %n = jmlBaris; d=jmlKolom  ..............1
   MVavg=nanmean(data); %untuk tipe numerik  ....................1
   MVmode=mode(data); %untuk tipe binominal atau ordinal  .......1
   hasil=data;           ........................................1
   for i = 1:n           ........................................2
      for j = 1:d        ........................................3
         if isnan(data(i,j))==1 %jika MV.........................4
            if j==2|j==3|j==6|j==7|j==9|j==11|j==12|j==13 %mode  ...5
               hasil(i,j)=MVmode(j); ..............................6
            else          ........................................6
               if j==10 %variabel ke-10 real ← mean  ..............7
                  hasil(i,j)=MVavg(j);  ...........................8
               else %round(mean)  .................................8
                  hasil(i,j)=round(MVavg(j));  ....................9
               end        ........................................9
            end           ........................................9
         end              ........................................9
      end                 ........................................9
   end                    ........................................9
   datasetMV=hasil;       ........................................10
end
```

Setelah UCR dan MVR, selanjutnya dilakukan SND yang ditangani dengan melakukan normalisasi (*Min-Max Normalization*) data dalam jangkauan 0-1. Berikut ini adalah fungsi SND yang dibuat menggunakan alat bantu Matlab:

```
function [datasetDN] = aaDN(data) %SND, data dari MVR
   [n, d]= size(data); %n = jmlBaris; d=jmlKolom  ..............1
   xMin=min(data,[],1);  ...................................1
   xMax=max(data,[],1);  ...................................1
   newMin=0;     ...........................................1
   newMax=1;     ...........................................1
   hasil=data;   ...........................................1
   for i = 1:n   ...........................................2
      for j = 1:d ..........................................3
         hasil(i,j)=((data(i,j)-xMin(j))/(xMax(j)-
xMin(j)))*((newMax-newMin)+newMin); ..........................4
      end        ..........................................4
   end           ..........................................4
   datasetDN=hasil;  .......................................5
end
```

Tabel 4.2 berikut ini menunjukkan hasil normalisasi data yang telah dilakukan menggunakan metode *Min-Max Normalization*.

Tabel 4.2 Hasil Normalisasi Data

No	X1	X2	X3 - X 12	X13	Y
1	0,71	1	…	0,75	0
2	0,80	1	…	0,00	1
…	…	…	…	…	…
1190	0,80	1	…	0,00	1

Setelah UCR, MVR, dan SND, selanjutnya dilakukan DV menggunakan metode *K-Fold Cross Validation* dengan parameter *K=10*. Metode ini dapat menjamin bahwa setiap *class* dan data terbagi merata untuk data latih dan data uji di setiap iterasi *K*. Metode lainnya yang juga mampu melakukan hal tersebut yaitu *Leave One Out*. Namun metode tersebut memberikan kompleksitas komputasi yang sangat besar, terlebih pada *dataset* yang berukuran besar. Oleh karena itu, metode ini menjadi pilihan yang lebih efisien.

4.1.2 Model K-NN dan Evaluasinya

Berikut ini merupakan fungsi K-NN yang dibuat menggunakan alat bantu Matlab:

```
function [outputKNN, lamaProses] = aaKNN(dataLatihInput,
dataLatihOutput, dataUjiInput, k, distance, voteRule)
  tic; ..................................................1
  outputKNN = knnclassify(dataUjiInput, dataLatihInput,
dataLatihOutput, k, distance, voteRule);.....................1
  lamaProses=toc; .......................................1
end
```

Berikut ini merupakan fungsi evaluasi menggunakan *Confusion Matrix* yang dibuat menggunakan alat bantu Matlab:

```
function [conMat, akurasi, kelasVSactual] =
aaEvaluasi(outputActual, outputModel)
  kelasVSactual = [outputModel outputActual]; ...............1
  conMat = confusionmat(outputActual, outputModel); .........1
  jmlDataTest = sum(conMat(:)); .............................1
  hasilBenar = sum(diag(conMat));  ..........................1
  akurasi = 100 * (hasilBenar / jmlDataTest);  ..............1
end
```

Berikut ini merupakan *script* pelatihan K-NN yang dibuat menggunakan alat bantu Matlab:

```
clc; clear; close all; warning off all;.....1
dsInput = xlsread('D:\aa Research\2019-2020, azis, alhamad, and
taliki, heart deases prediction using machine
learning\eksperimen\dsHeartDisease.xlsx','dsInput');.....1
dsOutput = xlsread('D:\aa Research\2019-2020, azis, alhamad, and
taliki, heart deases prediction using machine
learning\eksperimen\dsHeartDisease.xlsx','dsOutput');.....1
[datasetMV] = aaMVR(dsInput); % MVR.....1
[datasetDN] = aaDN(datasetMV); % SND.....1
dataset = [datasetDN, dsOutput];.....1
[m,n] = size(dataset); % m baris, n kolom.....1
subKNN=[];.....1
outputKNNall.x=[];.....1
myKNN = zeros(1,7);.....1
K=10;     .....1
indeks = crossvalind('Kfold', dataset(:,14), K);.....1
for i = 1:K.....2
  uji = (indeks == i); .....3
  latih = ~uji; .....3
  subDataLatihInput = dataset(latih,1:13); .....3
  subDataLatihOutput = dataset(latih,14); .....3
  subDataUjiInput = dataset(uji,1:13); .....3
  subDataUjiOutput = dataset(uji,14); .....3
  [outputKNN, lamaProses] = aaKNN(subDataLatihInput,
subDataLatihOutput, subDataUjiInput, 5, 'euclidean', 'nearest');
  subKNN(i,1)=lamaProses; .....3
  [conMat, akurasi, kelasVSactual] = aaEvaluasi(subDataUjiOutput,
outputKNN);.....3
  subKNN(i,2)=akurasi; .....3
  subKNN(i,3)=(conMat(1,1)/(conMat(1,1)+conMat(1,2)))*100; .....3
  subKNN(i,4)=(conMat(1,1)/(conMat(1,1)+conMat(2,1)))*100; .....3
  subKNN(i,5)=(conMat(2,2)/(conMat(2,2)+conMat(1,2)))*100; .....3
  outputKNNall(i).x = [subDataUjiOutput, outputKNN]; .....3
end       .....3
myKNN(1)=mean(subKNN(:,1));.....4
myKNN(2)=mean(subKNN(:,2));.....4
myKNN(3)=mean(subKNN(:,3));.....4
myKNN(4)=mean(subKNN(:,4));.....4
myKNN(5)=mean(subKNN(:,5));.....4
myKNN(6)=max(subKNN(:,2));.....4
myKNN(7)=min(subKNN(:,2));.....4
disp('Thanks to: Yolanda Djafar').....4
```

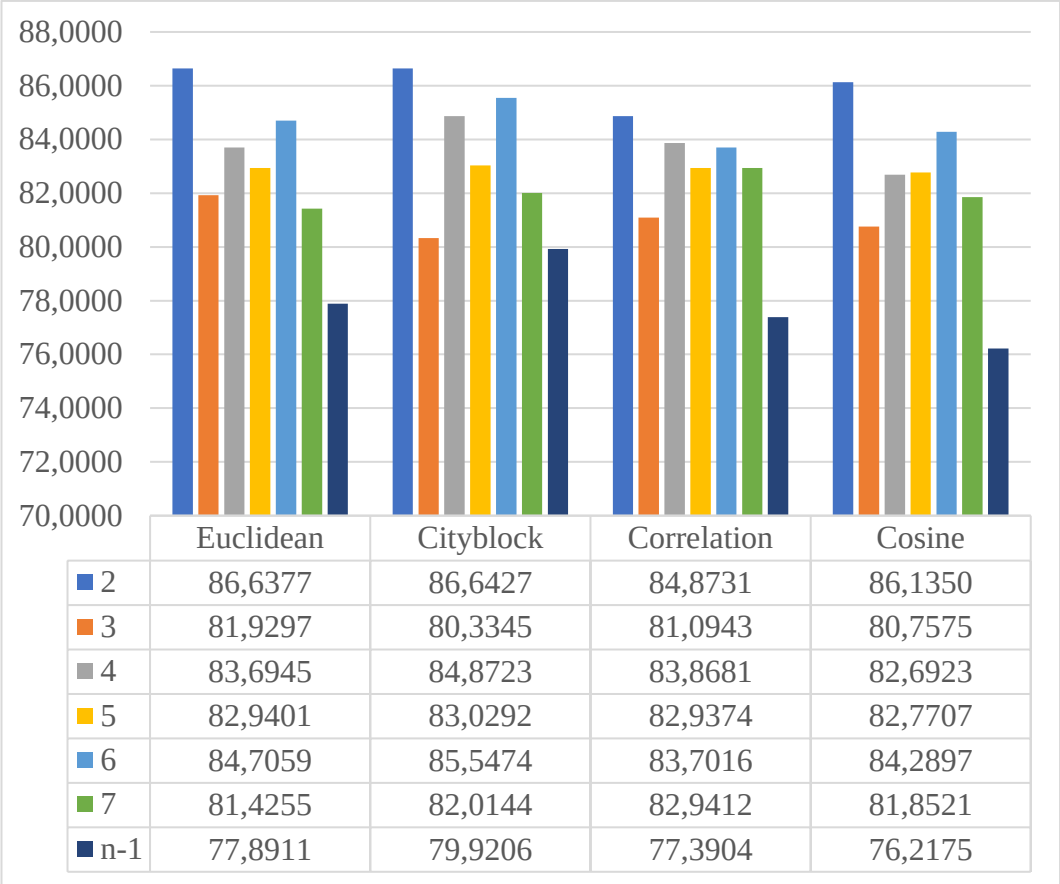
Tabel 4.3 berikut ini merupakan hasil evaluasi berbagai uji coba terhadap model-model K-NN berdasarkan Tabel 4.3 + UCR + MVR + SND

Tabel 4.3 Akurasi K-NN dengan Berbagai Parameter + UCR + MVR + SND

K	Distance	Vote	UCR	MVR	SND	Accuracy (%)
2	Euclidean	Nearest	Manually	Mean/Mode	Min-Max	86,6377
3	Euclidean	Nearest	Manually	Mean/Mode	Min-Max	81,9297
4	Euclidean	Nearest	Manually	Mean/Mode	Min-Max	83,6945
5	Euclidean	Nearest	Manually	Mean/Mode	Min-Max	82,9401
6	Euclidean	Nearest	Manually	Mean/Mode	Min-Max	84,7059
7	Euclidean	Nearest	Manually	Mean/Mode	Min-Max	81,4255
…	Euclidean	Nearest	Manually	Mean/Mode	Min-Max	…
n-1	Euclidean	Nearest	Manually	Mean/Mode	Min-Max	77,8911
2	**Cityblock**	**Nearest**	**Manually**	**Mean/Mode**	**Min-Max**	**86,6387**
3	Cityblock	Nearest	Manually	Mean/Mode	Min-Max	80,3345
4	Cityblock	Nearest	Manually	Mean/Mode	Min-Max	84,8723
5	Cityblock	Nearest	Manually	Mean/Mode	Min-Max	83,0292
6	Cityblock	Nearest	Manually	Mean/Mode	Min-Max	85,5474
7	Cityblock	Nearest	Manually	Mean/Mode	Min-Max	82,0144
…	Cityblock	Nearest	Manually	Mean/Mode	Min-Max	…
n-1	Cityblock	Nearest	Manually	Mean/Mode	Min-Max	79,9206
2	Correlation	Nearest	Manually	Mean/Mode	Min-Max	84,8731
3	Correlation	Nearest	Manually	Mean/Mode	Min-Max	81,0943
4	Correlation	Nearest	Manually	Mean/Mode	Min-Max	83,8681
5	Correlation	Nearest	Manually	Mean/Mode	Min-Max	82,9374
6	Correlation	Nearest	Manually	Mean/Mode	Min-Max	83,7016
7	Correlation	Nearest	Manually	Mean/Mode	Min-Max	82,9412
…	Correlation	Nearest	Manually	Mean/Mode	Min-Max	…
n-1	Correlation	Nearest	Manually	Mean/Mode	Min-Max	77,3904
2	Cosine	Nearest	Manually	Mean/Mode	Min-Max	86,1350
3	Cosine	Nearest	Manually	Mean/Mode	Min-Max	80,7575
4	Cosine	Nearest	Manually	Mean/Mode	Min-Max	82,6923
5	Cosine	Nearest	Manually	Mean/Mode	Min-Max	82,7707
6	Cosine	Nearest	Manually	Mean/Mode	Min-Max	84,2897
7	Cosine	Nearest	Manually	Mean/Mode	Min-Max	81,8521
…	Cosine	Nearest	Manually	Mean/Mode	Min-Max	…
n-1	Cosine	Nearest	Manually	Mean/Mode	Min-Max	76,2175

Dengan demikian, model K-NN yang terbaik untuk prediksi Penyakit Jantung adalah dengan *K* = 2, *Distance Measure = Cityblock, Vote Rule = Nearest* + UCR + MVR + SND, memberikan akurasi sebesar 86,6387%. Hasil uji coba terhadap parameter-parameter K-NN menunjukkan bahwa semakin tinggi nilai K,

maka semakin rendah pula akurasi. Secara grafik, hasil evaluasi ini ditunjukkan pada Gambar 4.1 berikut ini.



	Euclidean	Cityblock	Correlation	Cosine
2	86,6377	86,6427	84,8731	86,1350
3	81,9297	80,3345	81,0943	80,7575
4	83,6945	84,8723	83,8681	82,6923
5	82,9401	83,0292	82,9374	82,7707
6	84,7059	85,5474	83,7016	84,2897
7	81,4255	82,0144	82,9412	81,8521
n-1	77,8911	79,9206	77,3904	76,2175

Gambar 4.1 Akurasi K-NN dengan Berbagai Parameter + UCR + MVR + SND

Secara rinci, hasil evaluasi model K-NN yang diusulkan untuk prediksi Penyakit Jantung ini ditunjukkan pada Tabel 4.4 berikut ini.

Tabel 4.4 Hasil Evaluasi Model K-NN yang Diusulkan

Evaluation	**Result**
Accuracy	86,6427028438494%
Precision	87,3464912280702%
Specificity	85,0357062471124%
Sensitifity	88,5296647348006%
Maximum Accuracy	91,5966386554622%
Minimum Accuracy	83,1932773109244%
Times	0,00339902000000000 seconds

4.2 Hasil Pengembangan Sistem

4.2.1 Hasil Analisis Sistem

Gambar 4.2 Use Case Diagram

Gambar 4.3 Actifity Diagram

Gambar 4.4 Class Diagram

Gambar 4.5 Sequence Diagram

4.2.2 Hasil Desain Sistem

Gambar 4.6 Architecture Design (Model Jaringan Sistem)

Adapun spesifikasi hardware dan software yang direkomendasikan, yaitu:

1. Spesifikasi *Hardware*:
   - CPU : ?
   - GPU : ?
   - RAM : ?
   - Hardisk : ?
   - Resolusi : ?
2. Spesifikasi *Software*:
   - OS : ?
   - ?

Gambar 4.7 Mekanisme User

Gambar 4.8 Mekanisme Navigasi

Gambar 4.9 Mekanisme Input (Page)

Gambar 4.10 Mekanisme Output (Report)

Format data yang digunakan adalah *relational database* (SQL) menggunakan alat bantu My. SQL dengan struktur data sebagai berikut.

Tabel 4.5 Struktur Data Tabel User

Tabel 4.6 Struktur Data Tabel Latih

Tabel 4.7 Struktur Data Tabel Uji

Tabel 4.8 Struktur Data Tabel Average

Tabel 4.9 Struktur Data Tabel Hasil

Program Design dintujukkan pada Gambar 4.11 berikut ini.

Gambar 4.11 Program Design

4.2.3 Hasil Konstruksi Sistem

Pada tahap konstruksi sistem, hasil dari analisis dan desain sistem kemudian diterjemahkan ke konstruksi sistem/software menggunakan bahasa pemrograman Java dengan teknologi PHP. Adapun alat bantu yang digunakan pada tahap ini adalah ? untuk pemrogramannya, ? untuk database, dan ? untuk laporan-laporan.

4.2.4 Hasil Pengujian Sistem

Berikut ini merupakan hasil pengujian sistem menggunakan metode White Box, dimana program yang diuji adalah ?

Program ?

Gambar 4.12 Flowchart Program ? untuk Pengujian White Box

Gambar 4.13 Flowgraph Program ? untuk Pengujian White Box

Dengan demikian, perhitungan CC dari *flowgraph* tersebut (Gambar ?) pada pengujian White Box adalah sebagai berikut:

Diketahui:

Region (R) = ?

Node (N) = ?

Edge (E) = ?

Predikat Node (P) =?

Rumus: V(G) = (E – N) + 2 atau V(G) = P + 1

Penyelesaian:

V(G) = (24 – 17) + 2 = ?

V(G) = 8 + 1 = ?

(R1, R2, …, R?)

Tabel 4.10 Hasil Path Pengujian White Box

Setelah pengujian White Box, dilakukan pengujian Black Box yang ditunjukkan pada Tabel 4.11 berikut ini.

Tabel 4.11 Hasil Pengujian Black Box

BAB V
PEMBAHASAN

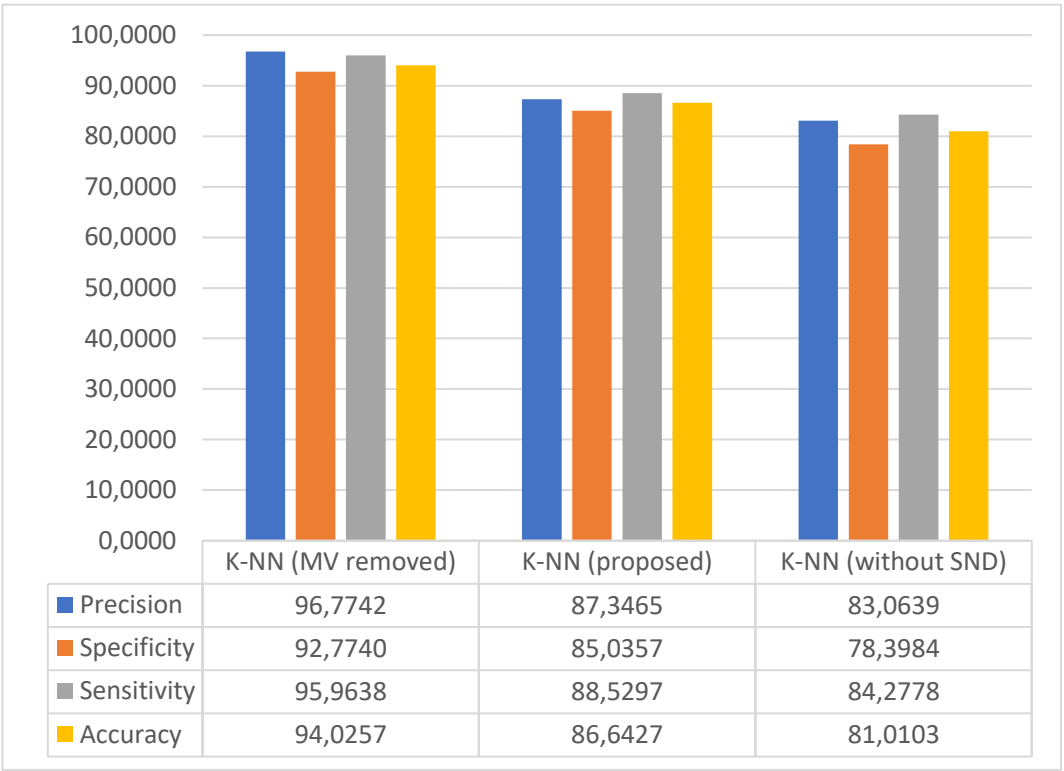
5.1 Pembahasan Model

Opsi komparasi antara model K-NN yang diusulkan dengan model K-NN lainnya ditunjukkan pada Tabel 5.1 berikut ini, di mana parameter K = 2, *Distance Measure* = *Cityblock*, dan *Vote Rule* = *Nearest* yang merupakan parameter terbaik K-NN untuk prediksi Penyakit Jantung.

Tabel 5.1 Opsi Komparasi

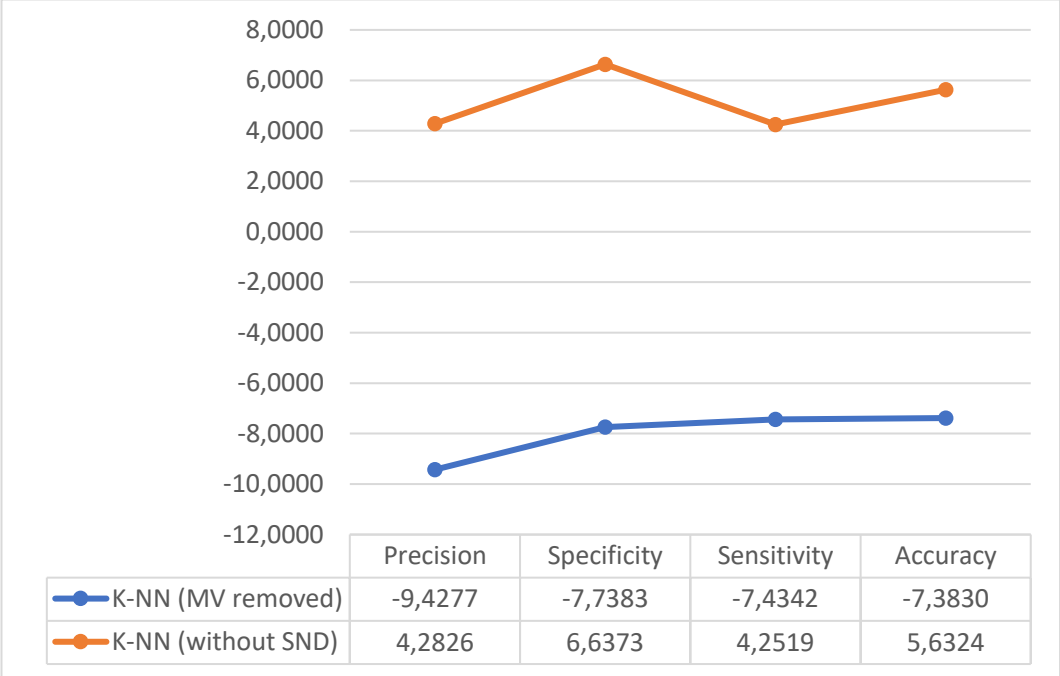
K-NN (MV removed)	**K-NN (proposed)**	**K-NN (without SND)**
UCR	UCR	UCR
No MVR (Removed)	MVR	MVR
No SND	SND	No SND

Gambar 5.1 berikut ini merupakan hasil komparasi antara K-NN (*MV removed*), K-NN (*proposed*), dan K-NN (without SND). Sedangkang Gambar 5.2 berikut ini merupakan selisih antara K-NN (proposed) dengan K-NN (*MV removed*) dan K-NN (without SND).



	K-NN (MV removed)	K-NN (proposed)	K-NN (without SND)
Precision	96,7742	87,3465	83,0639
Specificity	92,7740	85,0357	78,3984
Sensitivity	95,9638	88,5297	84,2778
Accuracy	94,0257	86,6427	81,0103

Gambar 5.1 Hasil Komprasi Kinerja K-NN



Gambar 5.2 Selisih Kinerja K-NN yang Diusulkan dengan Model K-NN Lainnya

Hasil komprasi ini mengindikasikan bahwa model K-NN yang diusulkan lebih akurat daripada model K-NN tanpat SND yaitu selisih 5,6324%, namun tidak lebih akurat daripada model K-NN dengan opsi MV *removed* (membuang MV) yaitu selisih -7,3830%. Walaupun dari sisi kinerja akurasi, K-NN yang diusulkan memang tidak lebih baik daripada K-NN (MV *removed*). Namun dari sisi efisiensi, kinerja K-NN yang diusulkan tentu saja lebih baik daripada K-NN (MV *removed*).

Opsi MV *removed* sebenarnya bukanlah merupakan solusi. Membuang MV yang begitu banyak, yaitu 1759 MV dari 1190 *instances* akan membuang sangat banyak *instances* yang mungkin saja memiliki informasi penting, tentu saja menghasilkan model/data yang *bias*, sehingga modelnya pun diragukan kinerjanya. Selain itu, membandingkan kinerja dua model dengan jumlah data yang berbeda tentu saja tidak adil, data yang berjumlah lebih sedikit cenderung akan lebih akurat. Opsi lainnya yang mungkin adalah memaksa K-NN agar dapat mengolah data yang memililki MV dengan pendekatan tertentu, namun opsi ini dapat menjadi penelitian ke depannya. Dengan demikian, dapat disimpulkan bahwa model K-NN yang diusulkan untuk prediksi Penyakit Jantung tersebut dapat meningkatkan kinerja akurasi dan efisiensi model K-NN untuk prediksi Penyakit Jantung.

5.2 Pembahasan Sistem

BAB VI
KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan, maka dapat disimpulkan bahwa model *K-Nearest Neighbor* (K-NN) yang diusulkan, yaitu dengan parameter *K* = 2, *Distance Measure = Cityblock*, *Vote Rule = Nearest*, *Unbalanced Class Reduction* secara manual, *Missing Value Replacement* menggunakan *mean/mode*, *Smothing Noisy Data* menggunakan *Min-Max Normalization*, dan *Data Validation* menggunakan *10-Fold Cross Validiation* memberikan akurasi sebesar 86,6427% lebih baik 5,6324% daripada tanpa *Smothing Noisy Data* dan lebih efisien daripada *Missing Value Removed*. Dengan demikian, model yang diusulkan dapat meningkatkan kinerja akurasi dan efisiensi model K-NN untuk prediksi Penyakit Jantung.

Walaupun dari sisi efisiensi, kinerja model K-NN yang diusulkan lebih baik daripada K-NN (*Missing Value Removed*), namun tidak lebih baik dari sisi akurasinya, selisih -7,3830%. Hal ini karena membuang 1759 *Missing Value* dari 1190 *instances* akan membuang sangat banyak *instances* yang mungkin memiliki informasi penting, sehingga justru menimbulkan *bias*, sehingga tidak adil untuk membandingkan akurasi dari kedua model tersebut dengan jumlah data yang berbeda jauh. Oleh karena itu, masalah ini dapat menjadi saran untuk penelitian selanjutnya, mungkin dengan menerapkan suatu pendekatan yang dapat memaksa K-NN mengolah data yang memilki *Missing Value*.

DAFTAR PUSTAKA

[1] Kementerian Kesehatan Republik Indonesia, "Penyakit Jantung Penyebab Kematian Tertinggi, Kemenkes Ingatkan CERDIK," 2017. [Online]. Available: http://www.depkes.go.id/article/view/17073100005/penyakit-jantung-penyebab-kematian-tertinggi-kemenkes-ingatkan-cerdik-.html. [Accessed: 19-Aug-2018].

[2] PT. Jawa Pos Grup Multimedia Redaksi, "Sepertiga Kematian di Dunia Dipicu Penyakit Jantung, Angkanya Segini," 2018. [Online]. Available: https://www.jawapos.com/kesehatan/29/09/2017/sepertiga-kematian-di-dunia-dipicu-penyakit-jantung-angkanya-segini. [Accessed: 19-Aug-2018].

[3] katadata, "Penyakit Kardiovaskular, Penyebab Kematian Terbanyak di Dunia," 2018. [Online]. Available: https://databoks.katadata.co.id/datapublish/2018/03/13/penyakit-kardiovaskular-penyebab-kematian-terbesar-di-dunia. [Accessed: 30-Aug-2018].

[4] S. H. Ishtake and S. A. Sanap, "Intelligent Heart Disease Prediction System Using Data Mining Techniques," *Int. J. Healthc. Biomed. Res.*, vol. 1, no. 3, pp. 94–101, 2013.

[5] M. Viceconti, P. Hunter, and R. Hose, "Big data, big knowledge: big data for personalized healthcare," in *IEEE Journal of Biomedical and Health Informatics*, 2015, vol. 19, no. 4, pp. 1209–1215.

[6] University of California Irvine Machine Learning Repository, "Heart Disease Dataset." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/heart+Disease.

[7] S. Bashir, U. Qamar, and F. H. Khan, "Heterogeneous classifiers fusion for dynamic breast cancer diagnosis using weighted vote based ensemble," *Qual. Quant.*, vol. 49, no. 5, pp. 2061–2076, 2015.

[8] I. Fakhruzi, "An Artificial Neural Network with Bagging to Address Imbalance Datasets on Clinical Prediction," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, no. 1, pp. 895–898.

[9] P. H. Abreu, M. S. Santos, M. H. Abreu, B. Andrade, and D. C. Silva, "Predicting Breast Cancer Recurrence Using Machine Learning Tehniques: A Systematic Review," *ACM Comput. Surv.*, vol. 49, no. 3, pp. 52:1-52:40, 2016.

[10] M. M. Suarez-Alvarez, D.-T. Pham, M. Y. Prostov, and Y. I. Prostov, "Statistical approach to normalization of feature vectors and clustering of mixed datasets," in *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 2012, vol. 468, no. 2145, pp. 2630–2651.

[11] G. E. A. P. A. Batista and M. C. Monard, “An analysis of four missing data treatment methods for supervised learning,” *Appl. Artif. Intell.*, vol. 17, no. 5–6, pp. 519–533, 2003.

[12] M. A. Jabbar and S. Shirina, “Heart disease prediction system based on hidden naïve bayes classifier,” in *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)*, 2016.

[13] S. Palaniappan and R. Awang, “Intelligent Heart Disease Prediction System using Data Mining Techniques,” *Int. J. Comput. Sci. Netw. Secur.*, vol. 8, no. 8, pp. 343–350, 2008.

[14] M. Anbarasi, E. Anupriya, and N. C. S. N. Iyengar, “Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm,” *Int. J. Eng. Sci. Technol.*, vol. 2, no. 10, pp. 5370–5376, 2010.

[15] J. Soni, U. Ansari, D. Sharma, and S. Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction,” *Int. J. Comput. Appl.*, vol. 17, no. 8, pp. 43–48, 2011.

[16] C. S. Dangare and S. S. Apte, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques,” *Int. J. Comput. Appl.*, vol. 47, no. 10, pp. 44–48, 2012.

[17] S. A. Pattekari and A. Parveen, “Prediction System for Heart Disease using Naive Bayes,” *Int. J. Adv. Comput. Math. Sci.*, vol. 3, no. 3, pp. 290–294, 2012.

[18] E. O. Olaniyi, O. K. Oyedotun, and A. Helwan, “Neural Network Diagnosis of Heart Disease,” in *International Conference on Advances in Biomedical Engineering (ICABME)*, 2015, pp. 21–24.

[19] J. Patel, T. Upadhyay, and S. Patel, “Heart Disease Prediction Using Machine Learning and Data Mining Technique,” *Comput. Sci. Electron. Journals*, vol. 7, pp. 129–137, 2016.

[20] J. Singh, A. Kamra, and H. Singh, “Prediction of Heart Diseases Using Associative Classification,” in *International Conference on Wireless Networks and Embedded Systems (WECON)*, 2016.

[21] R. El Bialy, M. A. Salama, and O. Karam, “An ensemble model for Heart disease data sets: a generalized model,” in *Proceedings of the 10th International Conference on Informatics and Systems - INFOS ’16*, 2016, pp. 191–196.

[22] Purushottam, K. Saxena, and R. Sharma, “Efficient Heart Disease Prediction System,” in *Procedia Computer Science*, 2016, vol. 85, pp. 962–969.

[23] S. Ekiz and P. Erdogmus, “Comparative Study of Heart Disease Classification,” in *Electric Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT)*, 2017.

[24] D. Kinge and S. K. Gaikwad, “Survey on Data Mining Techniques for

Disease Prediction," *Int. Res. J. Eng. Technol.*, vol. 5, no. 1, pp. 630–636, 2018.

[25] G. Manikandan, A. Vasudev, and A. Balasubramanian, "A Survey to Identify an Efficient Classification Algorithm for Heart Disease Prediction," *Int. J. Pure Appl. Math.*, vol. 119, no. 12, pp. 13337–13345, 2018.

[26] R. A. Kurian and K. S. Lakshmi, "An Ensemble Classifier for the Prediction of Heart Disease," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 3, no. 6, pp. 25–31, 2018.

[27] M. Shouman, T. Turner, and R. Stocker, "Using Data Mining Techniques in Heart Disease Diagnosis and Treatment," in *Japan-Egypt Conference on Electronics, Communications and Computers*, 2012, pp. 189–193.

[28] S. Vijayarani, "A Study of Heart Disease Prediction in Data Mining," *Int. J. Comput. Sci. Inf. Technol. Secur.*, vol. 2, no. 5, pp. 1041–1045, 2012.

[29] N. K. S. Banu and S. Swamy, "Prediction of Heart Disease at early stage using Data Mining and Big Data Analytics: A Survey," in *nternational Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*, 2016, pp. 256–261.

[30] B. Gnaneswar and M. R. E. Jebarani, "A Review on Prediction and Diagnosis of Heart Failure," in *International Conference on Innovations in Information, Embedded and Communication System (ICIIECS)*, 2017.

[31] C. Sowmiya and P. Sumitra, "Analytical Study of Heart Disease Diagnosis Using Classification Techniques," in *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, 2017.

[32] F. Gorunescu, *Data Mining: Concepts, Models, and Techniques*. Springer-Verlag Berlin Heidelberg, 2011.

[33] E. Prasetyo, *Data Mining, Mengolah Data Menjadi Informasi Menggunakan Matlab*. Andi Offset, 2014.

[34] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Elsevier, 2012.

[35] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*, 2nd ed. Wiley, 2014.

[36] S. Zhang, Z. Jin, and X. Zhu, "Missing data imputation by utilizing information within incomplete instances," *J. Syst. Softw.*, vol. 84, no. 3, pp. 452–459, 2011.

[37] M. S. Santos, P. H. Abreu, P. J. Garcia-Laencina, A. Simao, and A. Carvalho, "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients," *J. Biomed. Inform.*, vol. 58, pp. 49–59, 2015.

[38] University of California Irvine Machine Learning Repository, "Statlog

Dataset.” [Online]. Available: http://archive.ics.uci.edu/ml/datasets/statlog+(heart).

LAMPIRAN

Lampiran 1 Jadwal Penelitian

NO	KEGIATAN	2018		2019			
		11	12	1	2	3	4
1	Studi literatur						
2	Pengumpulan data						
3	Penyusunan proposal						
4	Pra pengolahan data						
5	Validasi data						
6	Eksperimen						
7	Evaluasi/pengujian						
8	Komparasi						
9	Hasil dan pembahasan						
10	Kesimpulan						
11	Penyusunan laporan						