PERBANDINGAN ALGORITMA NAIVE BAYES DAN DECISION TREE UNTUK KLASIFIKASI DATA REKAM MEDIS BERDASARKAN KODE ICD-10 DI RUMAH SAKIT X TANGERANG SELATAN

TUGAS AKHIR

Nabilah Mumtaz
41518010168

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2022

https://lib.mercubuana.ac.id/

PERBANDINGAN ALGORITMA NAIVE BAYES DAN DECISION TREE UNTUK KLASIFIKASI DATA REKAM MEDIS BERDASARKAN KODE ICD-10 DI RUMAH SAKIT X TANGERANG SELATAN

Tugas Akhir

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Sarjana Komputer

Oleh:
Nabilah Mumtaz
41518010168

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS ILMU KOMPUTER
UNIVERSITAS MERCU BUANA
JAKARTA
2022

https://lib.mercubuana.ac.id/

LEMBAR PERNYATAAN ORISINALITAS

Yang bertanda tangan dibawah ini:

NIM	:	41518010168
Nama	:	Nabilah Mumtaz
Judul Tugas Akhir	:	Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Klasifikasi Data Rekam Medis Berdasarkan Kode ICD-10 Di Rumah Sakit X Tangerang Selatan

Menyatakan bahwa Laporan Tugas Akhir saya adalah hasil karya sendiri dan bukan plagiat. Apabila ternyata ditemukan di dalam laporan Tugas Akhir saya terdapat unsur plagiat, maka saya siap untuk mendapatkan sanksi akademik yang terkait dengan hal tersebut.

Jakarta, 05 Februari 2022

METERAI TEMPEL

4DFCCAJX912537801

Nabilah Mumtaz

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id/

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa	:	Nabilah Mumtaz
NIM	:	41518010168
Judul Tugas Akhir	:	Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Klasifikasi Data Rekam Medis Berdasarkan Kode ICD-10 Di Rumah Sakit X Tangerang Selatan

Dengan ini memberikan izin dan menyetujui untuk memberikan kepada Universitas Mercu Buana **Hak Bebas Royalti Non Eksklusif** (*Non-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul di atas beserta perangkat yang ada (jika diperlukan).

Dengan Hak Bebas Royalti/Non Eksklusif ini Universitas Mercu Buana berhak menyimpan, mengalih media/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya.

Selain itu, demi pengembangan ilmu pengetahuan di lingkungan Universitas Mercu Buana, saya memberikan izin kepada Peneliti di Lab Riset Fakultas Ilmu Komputer, Universitas Mercu Buana untuk menggunakan dan mengembangkan hasil riset yang ada dalam tugas akhir untuk kepentingan riset dan publikasi selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 05 Februari 2022

UNIVERSITAS
MERCU BUANA

METERAI TEMPEL
GA29FAJX912537831

Nabilah Mumtaz

https://lib.mercubuana.ac.id/

SURAT PERNYATAAN LUARAN TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini

: Nama Mahasiswa	:	Nabilah Mumtaz
NIM	:	41518010168
Judul Tugas Akhir	:	Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Klasifikasi Data Rekam Medis Berdasarkan Kode ICD-10 Di Rumah Sakit X Tangerang Selatan

Menyatakan bahwa :

1. Luaran Tugas Akhir saya adalah sebagai berikut :

No	Luaran	Jenis		Status	
1	Publikasi Ilmiah	Jurnal Nasional Tidak Terakreditasi		Diajukan	
		Jurnal Nasional Terakreditasi			
		Jurnal Internasional Tidak Bereputasi		Diterima	
		Jurnal Internasional Bereputasi			
	Disubmit/dipublikasikan di :	Nama Jurnal	:		
		ISSN	:		
		Link Jurnal	:		
		Link File Jurnal Jika Sudah di Publish	:		

2. Bersedia untuk menyelesaikan seluruh proses publikasi artikel mulai dari submit, revisi artikel sampai dengan dinyatakan dapat diterbitkan pada jurnal yang dituju.
3. Diminta untuk melampirkan scan KTP dan Surat Pernyataan (Lihat Lampiran Dokumen HKI), untuk kepentingan pendaftaran HKI apabila diperlukan

Demikian pernyataan ini saya buat dengan sebenarnya.

Mengetahui
Dosen Pembimbing TA
Digitally signed by: SUKMA WARDHANA
Signed at: Aug 19, 2022 00:22:08

Sukma Wardhana
Sukma Wardhana S.Kom., M.Kom.

Jakarta, 06 Juli 2022

METERAI TEMPEL
F1BD3AJX912537806

Nabilah Mumtaz

UNIVERSITAS MERCU BUANA

https://lib.mercubuana.ac.id/

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518010168
Nama	:	Nabilah Mumtaz
Judul Tugas Akhir	:	Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Klasifikasi Data Rekam Medis Berdasarkan ICD-10 di Rumah Sakit X Tangerang Selatan

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 18 Agustus 2022

(Misabahul Fajri, M.TI)

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id/

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518010168
Nama	:	Nabilah Mumtaz
Judul Tugas Akhir	:	Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Klasifikasi Data Rekam Medis Berdasarkan ICD-10 di Rumah Sakit X Tangerang Selatan

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 18 Agustus 2022

(Rahmat Budiarto, Dr. Prof)

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id/

LEMBAR PERSETUJUAN PENGUJI

NIM	:	41518010168
Nama	:	Nabilah Mumtaz
Judul Tugas Akhir	:	Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Klasifikasi Data Rekam Medis Berdasarkan ICD-10 di Rumah Sakit X Tangerang Selatan

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 18 Agustus 2022

(Wawan Gunawan, S.Kom, MT)

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id/

LEMBAR PENGESAHAN

NIM	:	41518010168
Nama	:	Nabilah Mumtaz
Judul Tugas Akhir	:	Data Mining Rekam Medis untuk Menentukan Penyakit Terbanyak Menggunakan Decision Tree Dan Naive Bayes Berdasaarkan Kode ICD-10 Di Rumah Sakit X Tangerang Selatan

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 11 Agustus 2022

Menyetujui,

Sukma Wardhana

(Sukma Wardhana S. Kom., M. Kom.)
Dosen Pembimbing

Mengetahui,

(Wawan Gunawan, S.Kom., MT)	(Ir. Emil R. Kaburuan, Ph.D., IPM.)
Koord. Tugas Akhir Teknik Informatika	Ka. Prodi Teknik Informatika

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id/

KATA PENGANTAR

Puji syukur kita panjatkan kehadirat Allah SWT, tuhan yang Maha Esa, Maha pengasih lagi Maha Penyayang. Alhamdulillah atas segala rahmat-Nya sehingga penulis dapat menyelesaikan skripsi yang berjudul "Data Mining Rekam Medis untuk Menentukan Penyakit Terbanyak Menggunakan Decision Tree Dan Naive Bayes Berdasarkan Kode ICD-10 Di Rumah Sakit X Tangerang Selatan" dengan baik.

Dalam penyelesaian laporan Tugas Akhir ini, penulis tidak lepas dari dukungan, bantuan, serta sumbangan ide maupun pikiran juga materil dari berbagai pihak. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Bapak Sukma Wardhana S.Kom,M.Kom selaku Dosen Pembimbing Tugas Akhir yang memberikan Waktu, bimbingan dan masukan selama penyelesaian Tugas Akhir.
2. Bapak Emil Robert Kaburuan, PhD selaku Kepala Program Studi Teknik Informatika.
3. Bapak Wawan Gunawan , S.Kom, MT selaku Koordinator Tugas Akhir Jurusan Teknik Informatika.
4. Ibu Saruni Dwiasnati, ST, MM, M.Kom selaku Dosen pembimbing Akademik yang mengarahkan kegiatan akademik saya selama perkuliahan.
5. Kedua orangtua tercinta yang telah mendoakan, memberikan dukungan dan motivasi dalam menyelesaikan skripsi ini.
6. Bestie-bestie saya Adinda Siti Sholehah, Getsfy Mahardika, Astiana, NANASHERAPIT 5#, Rani Aprillya putri yang selalu memberikan semangat, keceriaan dan dukungan yang luar biasa sehingga dapat menyelesaikan skripsi ini.
7. Teman-teman dan semua pihak yang tidak dapat disebutkan namanya satu persatu yang telah membantu dan mau saya repotkan dalam penulisan skripsi ini.
8. Last but not least, I wanna thank me, I i thank me for believing in me, I wanna thank me for doing all this hard work, I wanna thank me for having

https://lib.mercubuana.ac.id/

no days off, I wanna thank me for never quitting, I wanna thank me for just being me at all times.

Akhir kata, penulis menyadari bahwa penyusunan Jurnal Tugas Akhir ini masih jauh dari kesempurnaan dan masih terdapat kekurangan dan kelemahan, walaupun demikian, penulis telah berusaha semaksimal mungkin untuk mendapatkan hasil yang terbaik. Akhir kata, penulis berharap tugas akhir ini dapat bermanfaat bagi pembaca guna menambah pengetahuan dan wawasan.

Jakarta, 20 Juni 2022
Nabilah Mumtaz

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id/

DAFTAR ISI

HALAMAN SAMPUL i
HALAMAN JUDUL ii
LEMBAR PERNYATAAN ORISINALITAS iii
SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR... iv
SURAT PERNYATAAN LUARAN TUGAS AKHIR v
LEMBAR PERSETUJUAN PENGUJI vi
LEMBAR PENGESAHAN ix
ABSTRAK x
ABSTRACT xi
KATA PENGANTAR xii
DAFTAR ISI xiv
NASKAH JURNAL 1
KERTAS KERJA 11
BAB 1. LITERATURE REVIEW 12
BAB 2. SOURCE CODE 24
BAB 3. DATASET 43
BAB 4. TAHAPAN EKSPERIMEN 44
BAB 5. HASIL SEMUA EKSPERIMEN 52
DAFTAR PUSTAKA 62
LAMPIRAN DOKUMEN HAKI 65
LAMPIRAN KORESPONDENSI 66

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id/

NASKAH JURNAL

Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Klasifikasi Data Rekam Medis Berdasarkan Kode ICD-10 Di Rumah Sakit X Tangerang Selatan

Comparison of Naive Bayes Algorithm and Decision Tree for Classification of Medical Record Data Based on ICD-10 Code at Hospital X South Tangerang

Nabilah Mumtaz[1], Sukma Wardhana[2]

[1, 2] Jurusan Teknik Informatika, Fakultas Ilmu komputer, Universitas Mercu Buana, Jakarta

[1]nabilahmumtaz79@gmail.com, [2]sukma@mercubuana.ac.id*

Abstract

The number of patient records that continues to increase can lead to data accumulation, impact on the performance of medical staff and to see trends in disease patterns suffered in community groups. The main objective of this research is to process medical data to classify the type of disease to make a positive contribution to all stakeholders, including hospital policy makers, medical institutions, and the government that handles health. The method used in this study is the Decision Tree and Naive Bayes algorithm method using the international disease code attribute as the destination label attribute for 21 international disease groups, namely: A00-B99 to Z00-Z99. From data mining obtained at Hospital X in South Tangerang for treatment in January 2019, data analysis was carried out. The accuracy value obtained by the decision tree algorithm is 0.93. In my opinion, this accuracy is already very high in predicting or classifying disease diagnoses at Hospital X in South Tangerang. And the accuracy value obtained by the Naive Bayes algorithm is only 0.61 which makes this algorithm have a final result that is not good enough because it produces a classification performance that is not high enough.

Keywords: Medical Records, Data Mining, Machine Learning, Decision Tree, Naive Bayes

Abstrak

Jumlah catatan pasien yang terus meningkat dapat menyebabkan akumulasi data, berdampak pada kinerja staf medis dan menyulitkan untuk melihat pola tren penyakit yang diderita dalam kelompok masyarakat. Tujuan utama dari penelitian ini adalah mengolah tumpukan data rekam medis untuk mengklasifikasikan jenis penyakit untuk memberikan kontribusi positif kepada semua pemangku kepentingan, baik pembuat kebijakan rumah sakit, institusi medis, dan pemerintah yang menangani kesehatan. Metode yang digunakan dalam penelitian ini adalah metode algoritma Decision Tree dan Naive Bayes dengan menggunakan atribut kode penyakit internasional sebagai atribut label tujuan sebanyak 21 kelompok penyakit internasional, yaitu: A00-B99 sampai dengan Z00-Z99. Dari data mining yang diperoleh di Rumah Sakit X di Tangerang Selatan berobat pada bulan Januari 2019 dilakukan analisis data. Nilai akurasi yang didapatkan oleh algoritma decision tree sebesar 0.93. nilai akurasi tersebut sudah sangat tinggi dalam memprediksi atau mengklasifikasikan diagnosa penyakit di Rumah Sakit X Tangerang Selatan. Dan Nilai akurasi yang didapatkan oleh algoritma naive bayes hanya sebesar 0.61 yang menjadikan algoritma ini memiliki hasil akhir yang tidak cukup bagus karena menghasilkan performa klasifikasi yang tidak cukup tinggi.

Kata kunci: Rekam Medis, *Data Mining*, *Machine Learning*, *Decision Tree*, Naive Bayes

https://lib.mercubuana.ac.id/

1. Pendahuluan

Pelayanan rawat jalan atau rawat inap merupakan salah satu unit fungsional rumah sakit yang menangani pasien rawat inap yang diterima dari pasien pada saat pendaftaran[1]. Informasi yang terkandung dalam rekam medis didasarkan pada tes, perawatan dan pengamatan. Rekam medis merupakan sumber informasi dan alat komunikasi yang dibutuhkan baik pasien maupun pasien. Penyedia layanan kesehatan dan penyedia layanan kesehatan harus dianggap sebagai penyedia layanan kesehatan untuk menetapkan kebijakan manajemen/kebijakan manajemen atau kebijakan perawatan kesehatan/manajemen. [2].

Berdasarkan informasi yang diterima dari Rumah Sakit Tangerang Selatan, Bagian Rekam Medis (RM) merupakan akumulasi dari berkas rekam medis, lancar dan cepat yang dapat mempengaruhi tenaga kesehatan. Di sisi lain, data rekam medis dapat diproses lebih lanjut untuk membantu profesional kesehatan memahami pola tren penyakit kelompok masyarakat dan untuk membantu profesional kesehatan memberikan pengobatan dan pengobatan untuk penyakit pasien[3]. Pengkodean ini dilakukan dengan menggunakan standar klasifikasi penyakit yang sesuai dengan ICD-10 atau *International Statistical Classification of Diseases and Related Health Problem Tenth Revision*[2][3]. Oleh karena itu, penulis ingin mengklasifikasikan data rekam medis berdasarkan *International Classification of Diseases ICD-10 (Diseases and Related Health Problems International Statistical Classification).*

Kode ICD ini berkembang terus, setiap waktu ada perubahan mendasar pada perkembangan penyakit baru yang ada di dunia. Sampai saat ini teknik pengkodean ICD tersebut mencapai ICD-10. ICD-10 adalah acuan seluruh penyelenggara pelayanan kesehatan (Rumah sakit, balai pengobatan, puskesmas) untuk dijadikan pedoman dalam melakukan arsip.

Pemilihan penggunaan metode algoritma serta data yang digunakan dalam penelitian ini adalah didasarkan dari penelitian-penelitian sebelumnya yang membahas penelitian dengan topik yang sama, yaitu pada penelitian yang dilakukan oleh Data Mining Rekam Medis Untuk Menentukan Penyakit Terbanyak Menggunakan Decision Tree C4.5[2], Penerapan Algoritma C4.5 untuk Klasifikasi Data Rekam Medis berdasarkan International Classification Diseases (ICD-10)[3], Perbandingan Klasifikasi Metode Naive Bayes dan Metode Decision Tree Algoritma (J48) pada Pasien Penderita Penyakit Stroke di RSUD Abdul Wahab Sjahranie Samarinda[4]

Dari data yang sangat besar tersebut, kemudian dapat dirumuskan permasalahan nya sebagai berikut:

1. Bagaimana hasil ketepatan klasifikasi data mining rekam medis untuk menentukan penyakit terbanyak menggunakan Decision Tree Dan Naive Bayes berdasarkan Kode ICD-10 di Rumah Sakit X Tangerang Selatan?
2. Bagaimana hasil ketepatan klasifikasi data mining rekam medis untuk menentukan penyakit terbanyak menggunakan Decision Tree Dan Naive Bayes berdasarkan Kode ICD-10 di Rumah Sakit X Tangerang Selatan?
3. Bagaimana perbandingan ketepatan hasil klasifikasi data mining rekam medis untuk menentukan penyakit terbanyak menggunakan Decision Tree Dan Naive Bayes berdasarkan Kode ICD-10 di Rumah Sakit X Tangerang Selatan?

Diharapkan pada penelitian ini dapat dijadikan sumber informasi untuk pihak yang dituju. Hal ini bertujuan supaya mempermudah di beberapa unit rumah sakit seperti IGD dan Poliklinik dalam mendiagnosa penyakit pasien.

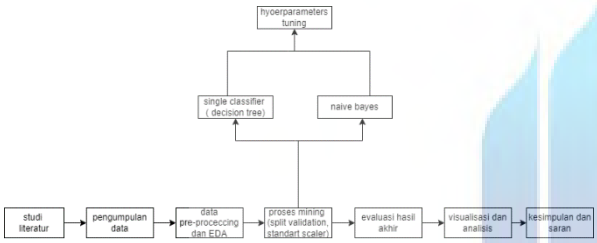
https://lib.mercubuana.ac.id/

2. Metode Penelitian

Penelitian fokus pada komparasi dan akurasi menganalisis data rekam medis menggunakan Decision Tree dan Naive Bayes. Untuk memudahkan dalam penerapan metodologi dan perancangan sistem maka dibuat bagan alir analisa dan perancangan seperti dibawah ini:

A. Alur Penelitian

Dalam penelitian ini terdiri dari beberapa tahapan proses. Dimulai dari studi literatur, pengumpulan data, data preprocessing dan EDA, pra-pemodelan (split validation, standard scaler), pemodelan dengan algoritma machine learning, tuning hyperparameters kedua algoritma, validasi hasil akhir, analisis hasil akhir, selesai. *Tahapan penelitian bisa dilihat pada gambar 1.*



Gambar 1.

B. Rincian Alur Penelitian

Alur kerja Penelitian ini ditunjukkan pada gambar 2, penjelasan setiap tahapannya sebagai berikut:

1. Studi Literatur
   Pada tahap ini, penulis melakukan tinjauan pustaka dari berbagai referensi yang terkait dengan penelitian yang dilakukan. Dalam penelitian ini, penulis mengkaji tahapan pengumpulan atau ekstraksi data, konsep data mining. Implementasi algoritma klasifikasi dan hasil penyakit terbanyak di Rumah Sakit X Tangerang Selatan.
2. Pengumpulan Data
   Dalam penelitian ini adalah dataset diambil dari Rumah Sakit X Tangerang Selatan. Data yang diambil yaitu dataset rekam medis yang didalamnya berkaitan dengan diagnosa penyakit berdasarkan kode ICD-10. Data yang terkumpul ± sebanyak 3.178 data rekam medis.
3. Pra-Processing dan EDA
   Pengolahan data (preprocessing) merupakan tahap pertama untuk melakukan proses data mining yang terdiri dari pembersihan dan transformasi data. Preprocessing ini dimaksudkan untuk mentransformasikan data yang digunakan pada aplikasi data mining agar mudah diinterpretasikan untuk dianalisis. Setelah preprocessing, EDA dijalankan. Setelah pra pemroses memahami kumpulan data yang dibuat menggunakan alat statistik dasar, merencanakan dan memvisualisasikan data sebelum memulai pembelajaran machine learning.
4. Proses Mining (Pemodelan dan Pengujian)

Pemodelan & Hyperparameter Tuning

Selanjutnya dilakukan proses mining dan hyperparameter tuning untuk mendapatkan hasil pengujian yang lebih optimal dengan menggunakan algoritma Decision Tree dan Naive Bayes. Machine Learning adalah jenis kecerdasan buatan (AI) yang memungkinkan Anda memprediksi hasil dengan lebih akurat, bahkan jika aplikasi perangkat lunak Anda tidak diprogram secara eksplisit[5]. pembelajaran mesin learning menggunakan data historis sebagai input untuk memprediksi nilai output baru.

a) Decision Tree

Decision tree adalah prediksi terhadap suatu keputusan menggunakan struktur hirarki atau pohon. Setiap pohon memiliki cabang yang mewakili setiap atribut yang harus dipenuhi untuk menuju cabang selanjutnya hingga tidak ada cabang lagi. Dalam pembuatan decision tree, terdapat nilai entropy dan gain, dengan rumus sebagai berikut:

- menghitung nilai entropy digunakan rumus:

$$Entropy(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i$$

- S = himpunan kasus
- n = jumlah partisi S
- pi = proporsi Si terhadap S

- menghitung nilai gain digunakan rumus:

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Valus(A)} \frac{|S_v|}{|S|} Entropy\ (S_v)$$

- A : Atribut

UNIVERSITAS MERCU BUANA

- S : Sampel
- n : Jumlah partisi himpunan atribut A
- |Si| : Jumlah sampel pada partisi ke –i
- |S| : Jumlah sampel dalam S

b) Naive Bayes

Naive Bayes Classifier ini menerapkan teknik *supervised* klasifikasi objek di masa depan dengan menetapkan label kelas ke *instance*/catatan menggunakan probabilitas bersyarat. Probabilitas bersyarat adalah ukuran peluang suatu peristiwa yang terjadi berdasarkan peristiwa lain yang telah (dengan asumsi, praduga, pernyataan, atau terbukti) terjadi. model *machine learning* yang diterapkan pada program tersebut menggunakan teorema Bayes yang dirumuskan sebagai berikut:

$$P(A \mid B) = P(B \mid A)P(A)P(B)$$

Keterangan:

***P(A | B)* : Probabilitas A terjadi dengan bukti bahwa B telah terjadi (probabilitas superior)**

P(B | A) : Probabilitas B terjadi dengan bukti bahwa A telah terjadi

P(A) : Peluang terjadinya A

P(B) : Peluang terjadinya B

c) Cross Validation

Tujuan dari validasi silang adalah untuk menguji kemampuan model untuk memprediksi data baru yang tidak akan digunakan untuk estimasi, menandai masalah seperti *overfitting* dan bias seleksi, dan membuat model menjadi kumpulan data independen (yaitu, kumpulan data yang tidak diketahui, misalnya, tidak diketahui dataset dari masalah sebenarnya)[6].

- K fold Cross Validation

Cross validation adalah suatu metode tambahan dari teknik data mining yang bertujuan untuk memperoleh hasil akurasi yang maksimal. Metode ini sering juga disebut dengan *k-fold cross validation* dimana percobaan sebanyak *k* kali untuk satu model dengan parameter yang sama Dalam bukunya yang berjudul "*Data Mining dan Big Data Analytics*"[7].

d) Hyperparameter tuning

Hyperparameter Tuning adalah nilai untuk parameter yang digunakan untuk mempengaruhi proses pembelajaran[8]. *Hyperparameter tuning* adalah proses mengidentifikasi *tuple hyperparameters* yang menghasilkan model optimal yang meminimalkan fungsi kerugian yang telah diberikan data independen yang disediakan.

- GridSearchCV

GridSearchCV merupakan bagian dari modul *scikit-learn* yang bertujuan untuk memvalidasi secara otomatis dan sistematis beberapa model dan setiap hyperparameter [9].

Pengujian

Pengujian dilakukan menggunakan data test dengan variabel independen ada kolom jenis jaminan, umur pasien, dokter. Lalu di variabel dependen ada kolom ruangan. Kolom diagnosa tidak digunakan pada variabel dependen atau kolom prediksi karena nilai unik di dalam kolom tersebut terlalu banyak sehingga membuat akurasi keseluruhan sera nilai presisi, recall, dll. antar nilai unik tersebut sangat kecil. *Evaluation metrics* yang paling umum digunakan untuk model klasifikasi yaitu akurasi, presisi, recall dan f1-score.

5. Evaluasi Proses Mining

   Fase ini merupakan evaluasi terhadap algoritma yang digunakan pada fase ini. Menghasilkan hasil evaluasi dengan membandingkan nilai akurasi yang diperoleh untuk menemukan hasil akurasi yang paling efektif.

6. Visualisasi dan Analisis

   Pada tahapan ini dilakukan analisis dan visualisasi dataset serta hasil implementasi pengklasifikasian dengan data mining untuk mengetahui perbandingan dari kedua algoritma serta hasil klasifikasi penyakit terbanyak di rumah sakit x tangerang selatan.

7. Kesimpulan dan Saran

   Setelah proses analisis dan visualisasi tahapan terakhir dalam penelitian ini adalah menarik kesimpulan dari apa yang sudah penulis lakukan dalam penelitian ini berupa algoritma yang paling efektif dalam pengklasifikasian dataset untuk menentukan

https://lib.mercubuana.ac.id/

penyakit terbanyak dan analisis data untuk perguruan tinggi serta saran untuk penelitian selanjutnya.

3. Hasil dan Pembahasan

Dari tahapan penelitian diatas didapatkan hasil penelitian ini adalah sebagai berikut:

A. Hasil Pre-Processing

	Tanggal	Ruangan	Nama	Jaminan Kesehatan	Dokter	Diagnosa	Umur Pasien
0	31-01-2019	4	Ny. ETI SURYATI	0	1	377	8
1	31-01-2019	4	Ny. ANITA SELVILLA R	0	1	330	21
2	31-01-2019	4	Tn. IR H.ARIFIN EFENDI	0	1	377	31
3	31-01-2019	4	Ny. YULIAH	0	1	479	30
4	31-01-2019	4	Ny. MARDIANI SENOKEY	0	1	638	13
5	31-01-2019	4	Ny. RATNA MULAYAN	0	1	377	13
6	31-01-2019	4	Ny. ENA YURNALIS	0	1	348	27
7	31-01-2019	4	Ny. LINDA YANTI	0	1	380	14
8	31-01-2019	4	Ny. KARSIH	0	1	42	5
9	31-01-2019	4	Ny. SOFNIARTY	0	1	478	19
10	31-01-2019	4	Ny. BERLIANA TAMPUBOLON	0	1	377	19
11	31-01-2019	4	Ny. LIM ME ING	0	1	357	7

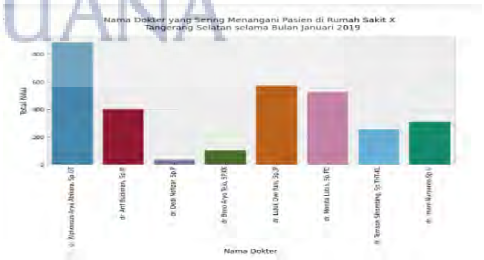
Gambar 2.

Berdasarkan hasil output diatas dapat diketahui bahwa keempat kolom di dalam Data Frame "data_diagnosa" telah memiliki tipe data yang berbeda yaitu integer sehingga untuk saat ini tipe data yang ada di "data_diagnosa" memiliki 2 tipe data yang berbeda yaitu integer dan object/string. Namun, jika diperhatikan kembali kolom "Tanggal" seharusnya memiliki tipe data datetime bukan object. Maka dari itu untuk proses selanjutnya saya akan melakukan perubahan tipe data yang ada di kolom "Tanggal" dan "Umur Pasien" dari yang semula object/string menjadi date time untuk kolom "Tanggal" dan integer untuk kolom "Umur Pasien, sesuai dengan tujuan/maksud dari kolom tersebut yaitu menunjukkan entitas waktu terhadap tanggal masuknya pasien ke rumah sakit x di Tangerang Selatan serta menunjukkan umur pasien yang bersifat angka numerik.

B. Exploratory Data Analysis (EDA) with Graphs

Gambar 3.

Berdasarkan grafik batang di atas telah diketahui bahwa ruangan yang paling banyak digunakan di rumah sakit x Tangerang Selatan adalah ruangan ortopedi, ruangan jantung dan ruangan penyakit dalam. Sebaliknya, ruangan yang jarang digunakan di rumah sakit x Tangerang Selatan adalah ruangan paru, odc, serta ruangan kulit dan kelamin. Atribut "Ruangan" yang ada di DataFrame, baik yang asli ("data_diagnosa") ataupun hasil copy ("data_diagnosa_copy"), menurut saya sangat berkaitan dengan atribut "Diagnosa" karena pasien yang ada di ruangan-ruangan ini merupakan hasil diagnosa yang telah dilakukan oleh para dokter yang menangani penyakit-penyakit ini. Maka dari itu dapat disimpulkan bahwa ruangan ortopedi, jantung dan penyakit merupakan ruangan yang memiliki tingkat diagnosa tertinggi sekaligus terbanyak dibandingkan dengan ruangan-ruangan lain yang ada di rumah sakit x Tangerang Selatan.

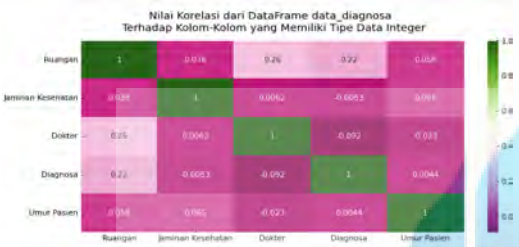


Gambar 4.

Grafik *gambar 4* menunjukkan nama-nama dokter yang sering menangani pasien di bulan Januari tahun 2019 di rumah sakit X Tangerang Selatan sehingga dapat disimpulkan bahwa dr. Alphonsus Arya Abikara, Sp.OT merupakan dokter yang sering mendapatkan pasien di rumah sakit x. Selain dokter Alphonsus Arya Abikara, ada juga dokter yang bernama dr. Luluk Dwi Yuni, Sp.JP dan dr. Monita Lubis, Sp.PD yang juga sering mendapatkan serta

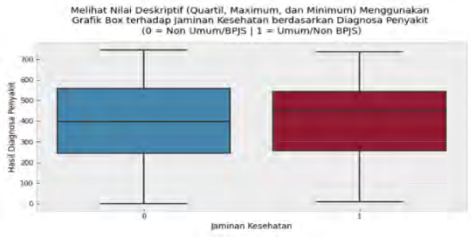
https://lib.mercubuana.ac.id/

mendiagnosa pasien yang berkunjung dalam periode bulan Januari tahun 2019. Dokter Alphonsus Arya Abikara telah mendiagnosa pasien sebanyak 914 pasien rumah sakit x. Dokter Luluk Dwi Yuni, Sp.JP telah mendiagnosa pasien rumah sakit x sebanyak 576 serta dokter Monita Lubis, Sp.PD yang juga telah mendiagnosa pasien sebanyak 553 pasien di rumah sakit x Tangerang Selatan. Sebaliknya, dokter yang jarang mendiagnosa pasien rumah sakit x atau bisa dikatakan sebagai dokter yang jarang menerima pasien adalah dokter yang bernama dr. Dedi Novizar. Sp.P yang telah mendiagnosa pasien rumah sakit x sebanyak 32 saja.



Gambar 5.

Grafik heatmap bertujuan untuk memberikan informasi terkait seberapa tinggi suatu atribut memiliki hubungan timbal-balik atau nilai korelasi dengan atribut yang lain berdasarkan nilai matriks yang ada didalamnya. Intinya setiap kolom/atribut yang ada di SUMBU Y (yang berada di sisi kiri) dibaca setiap nilainya dengan kolom-kolom yang berada di SUMBU X (yang berada di bawah). Untuk memudahkan dalam membedakan apakah nilai korelasinya bagus atau tidak dapat dilihat dari indikator warna yang di sebelah paling kanan dari grafik heat map/peta panas. Semakin bagus nilai korelasinya maka warnanya menjadi hijau tua tetapi semakin buruk nilai korelasinya maka nilai korelasinya akan menjadi berwarna ungu gelap seperti yang diperlihatkan antara atribut "Dokter" dengan "Diagnosa".



Gambar 6.

Grafik box merupakan sebuah grafik yang dapat menggambarkan bentuk distribusi suatu kolom data dengan menampilkan kelima unsur nilai statistik deskriptif dari kolom tersebut diantaranya nilai maksimum/nilai observasi tertinggi, nilai kuartil 3 (0.75), nilai kuartil 2 atau juga dapat dikatakan sebagai nilai median (0.50), nilai kuartil 1 (0.25), dan nilai minimum/nilai observasi terendah. Di dalam grafik box juga dapat diketahui apakah kolom yang digunakan sebagai label y atau variabel y memiliki nilai outlier atau tidak karena nilai outlier tersebut dapat mempengaruhi bentuk grafik yang diperlihatkan oleh kolom yang digunakan sebagai label x atau variabel x. Nilai outliers tersebut biasanya ditunjukkan diatas nilai maksimum atau dibawah nilai minimum. Berikut adalah penjelasan dari grafik box berdasarkan hasil output diatas berdasarkan kolom "Jaminan Kesehatan" dan "Diagnosa":

1. Jaminan Kesehatan dengan Nilai 0 (Non Umum/BPJS):
   - Dari grafik box terhadap jaminan kesehatan dengan nilai 0 tidak memiliki nilai outliers, baik dari hasil observasi nilai tertinggi (nilai maksimum) maupun observasi nilai terendah (nilai minimum).
   - Nilai maximum terhadap grafik jaminan kesehatan 0 adalah 745.
   - Nilai kuartil 3 terhadap grafik jaminan kesehatan dengan nilai 0/non umum berada di range nilai kuartil sebesar 558.00.
   - Nilai kuartil 2/nilai median dari grafik jaminan kesehatan umum adalah sebesar 398.0.
   - Nilai kuartil 1 terhadap grafik jaminan kesehatan non umum yang ditandai dengan nilai 0

https://lib.mercubuana.ac.id/

berada di range nilai kuartil sebesar 245.25.

- Terakhir adalah nilai minimum dari grafik non umum atau yang menggunakan jaminan kesehatan BPJS menunjukkan nilai 0. Sedangkan nilai interquartile range untuk nilai 0 ini adalah 3 (53 (kuartil 3) - 50 (kuartil 1) = 3).

2. Jaminan Kesehatan dengan Nilai 1 (Umum/Non BPJS):
   - Dari grafik box terhadap jaminan kesehatan dengan nilai 1 tidak memiliki nilai outliers, baik dari hasil observasi nilai tertinggi (nilai maksimum) maupun observasi nilai terendah (nilai minimum).
   - Nilai maximum terhadap grafik jaminan kesehatan 1 adalah 737.
   - Nilai kuartil 3 terhadap grafik jaminan kesehatan dengan nilai 1/umum berada di range nilai kuartil sebesar 542.25.
   - Nilai kuartil 2/nilai median dari grafik jaminan kesehatan umum adalah sebesar 450.0.
   - Nilai kuartil 1 terhadap grafik jaminan kesehatan umum yang ditandai dengan nilai 1 berada di range nilai kuartil sebesar 257.25.
   - Terakhir adalah nilai minimum dari grafik umum yang berada di range nilai 10.

Gambar 7.

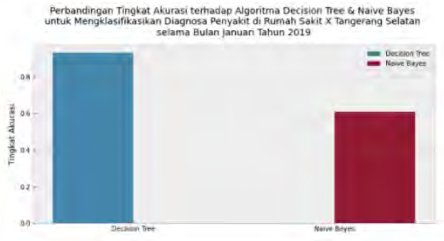
Berdasarkan grafik time series diatas telah diketahui karakteristik terkait waktu keluarnya hasil diagnosa untuk para pasien rumah sakit x Tangerang Selatan. Maka dari itu dapat disimpulkan bahwa dari tanggal 2 Januari 2019 sampai 31 Januari 2019, trend terhadap pemeriksaan serta hasil pemeriksaan berupa diagnosa penyakit yang ditujukan kepada para pasien didominasi dengan peningkatan daripada penurunan bahkan bentuk sifat dari grafik garis di atas lebih menonjol unsur peningkatan daripada penurunan trend. Peningkatan trend terhadap diagnosa penyakit tersebut mungkin saja dipengaruhi oleh jumlah kedatangan pasien-pasien di tanggal tersebut sekaligus hasil diagnosa sehingga mengakibatkan kenaikan trend di beberapa periode tanggal. Sebaliknya, penurunan trend terhadap hasil diagnosa penyakit yang ditujukan kepada pasien rumah sakit x mungkin disebabkan karena faktor tidak padat kunjungan pasien di rumah sakit untuk diperiksa atau sekadar kontrol sehingga mengakibatkan trend diagnosa di tanggal-tanggal tertentu mengalami penurunan seperti yang ditunjukkan di tanggal 21 Januari 2019.

Gambar 8.

Berdasarkan grafik batang diatas dapat diketahui bahwa pasien yang berumur 11 tahun memiliki jumlah total yang cukup dominan, yang ditunjukkan oleh entitas batang (bar) di umur 11 tahun, sehingga dapat dikatakan bahwa pasien yang memiliki umur 11 tahun adalah pasien terbanyak di rumah sakit x Tangerang Selatan yang mendapatkan diagnosa dari dokter. Jumlah nilai dari Umur 11 tahun sebanyak 144 data. Sebaliknya, umur pasien yang paling sedikit atau tidak mendominasi di rumah sakit x Tangerang Selatan adalah umur 49 tahun karena hanya memiliki total data yaitu 1 saja.

C. Membuat Grafik Hasil Akhir untuk Perbandingan Tingkat Akurasi Antara Algoritma Decision Tree dan Naive Bayes

https://lib.mercubuana.ac.id/



Gambar 9.

Berdasarkan grafik batang diatas yang menunjukkan tingkat perbedaan antara akurasi decision tree dan akurasi naive bayes.

4. Kesimpulan

Berdasarkan hasil eksperimen yang telah dilakukan, penelitian ini berhasil menganalisis kinerja algoritma Decision Tree dan Naive Bayes terhadap data rekam medis untuk menentukan penyakit terbanyak. Dari kesimpulan yang bisa diambil :

1. Nilai akurasi yang didapatkan oleh algoritma decision tree sebesar 0.93. Menurut saya akurasi tersebut sudah sangat tinggi dalam memprediksi atau mengklasifikasikan diagnosa penyakit di rumah sakit x Tangerang Selatan. Selain itu, algoritma decision tree dapat digunakan untuk mendukung sebuah pengambilan keputusan dengan tepat. Maka dari itu algoritma decision tree sangat disarankan untuk digunakan kembali dalam Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Klasifikasi Data Rekam Medis Berdasarkan Kode ICD-10 Di Rumah Sakit X Tangerang Selatan.

2. Nilai akurasi yang didapatkan oleh algoritma naive bayes hanya sebesar 0.61 yang menjadikan algoritma ini memiliki hasil akhir yang tidak cukup bagus karena menghasilkan performa klasifikasi yang tidak cukup tinggi. Maka dari itu, menurut asumsi saya, algoritma naive bayes tidak disarankan untuk menjadi pendukung keputusan dalam Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Klasifikasi Data Rekam Medis Berdasarkan Kode ICD-10 Di Rumah Sakit X Tangerang Selatan menggunakan data-data yang ada di masa depan.

UNIVERSITAS
MERCU BUANA

Daftar Rujukan

[1] D. Rusliyanti, “Analisis Ketepatan Pengkodean Diagnosis Berdasarkan ICD-10 dengan Penerapan Karakter Ke-5 pada Pasien Fraktur Rawat Jalan Semester II di RSU Mitra Paramedika Yogyakarta,” *J. Permata Indones.*, vol. 7, no. 1, pp. 26–34, 2016, [Online]. Available:

https://lib.mercubuana.ac.id/

http://www.permataindonesia.ac.id/wp-content/uploads/2016/08/03.-Jurnal-PI_Lusi-Anas-Harinto.pdf.

[2] Y. Mardi, "Data Mining Rekam Medis Untuk Menentukan Penyakit Terbanyak Menggunakan Decision Tree C4.5," *J. Sains dan Inform.*, vol. 4, no. 1, pp. 40–53, 2018, doi: 10.22216/jsi.v4i1.3077.

[3] Y. A. Fiandra, S. Defit, and Y. Yuhandri, "Penerapan Algoritma C4.5 untuk Klasifikasi Data Rekam Medis berdasarkan International Classification Diseases (ICD-10)," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 1, no. 2, pp. 82–89, 2017, doi: 10.29207/resti.v1i2.48.

[4] I. Lishania, R. Goejantoro, and Y. N. Nasution, "Perbandingan Klasifikasi Metode Naive Bayes dan Metode Decision Tree Algoritma (J48) pada Pasien Penderita Penyakit Stroke di RSUD Abdul Wahab Sjahranie Samarinda," *J. Eksponensial*, vol. 10, no. 2, pp. 135–142, 2019, [Online]. Available: http://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/view/571.

[5] Robby Takdirillah, "Apa itu Machine Learning? Beserta Pengertian dan Cara Kerjanya," *blog dicoding*, 2020. https://www.dicoding.com/blog/machine-learning-adalah/ (accessed Aug. 19, 2022).

[6] P. M. Cross-validation, "Cross-validation (statistics)," pp. 1–15.

[7] Abdul Muiz Khalimi, "Pengujian Data dengan Cross Validation," *pengalaman-edukasi blog*, 2020. https://www.pengalaman-edukasi.com/2020/04/apa-itu-k-fold-cross-validation.html.

[8] Adella Gravita, "Kenali Hyperparameter Tuning dalam Machine Learning," *blog Coding Studio*. https://codingstudio.id/hyperparameter-tuning/.

[9] MUHAMMAD ARIQ LEESTA HIDAYAT, "GRIDSEARCHCV," *algotech blog*, 2021. https://algotech.netlify.app/blog/gridsearchcv/#:~:text=GridSearchCV merupakan bagian dari modul,masing secara otomatis dan sistematis.

[10] Z. A. Amin, W. Cholil, M. I. Herdiansyah, and E. S. Negara, "Analisa Rekam Medis Elektronik Untuk Menentukan Diagnosa Medis Dalam Kategori Bab ICD 10 Menggunakan Machine Learning," *POSITIF J. Sist. dan Teknol. Inf.*, vol. 7, no. 2, pp. 127–132, 2021, doi: 10.31961/positif.v7i2.1140.

[11] J. Wandana, S. Defit, and S. Sumijan, "Klasterisasi Data Rekam Medis Pasien Pengguna Layanan BPJS Kesehatan Menggunakan Metode K-Means," *J. Inf. dan Teknol.*, vol. 2, pp. 4–9, 2020, doi: 10.37034/jidt.v2i4.73.

[12] B. Untuk, K. Data, and R. Medis, "1 , 2 , 3," vol. 10, pp. 407–415, 2021.

[13] A. A. PUTRI;, "ANALISIS KEAKURATAN KODE DIAGNOSIS PENYAKIT BERDASARKAN ICD-10 DALAM REKAM MEDIS RAWAT JALAN DI PUSKESMAS TLOGOSARI KULON KOTA SEMARANG," 2018, Accessed: Jun. 11, 2022. [Online]. Available: http://repository.poltekkes-smg.ac.id/index.php?p=show_detail&id=16023&keywords=.

[14] E. Darnila, H. Al-kautsar, and Y. Iksan, "Analysis of Patient Disease Trends Based on Medical Record Data Using the C4 . 5 Algorithm," vol. 14, no. 1, pp. 7–12, 2020.

[15] S. Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti, N. Chamidah, M. Mega Santoni, N. Matondang, F. Ilmu Komputer, and U. Pembangunan Nasional Veteran Jakarta, "Terakreditasi SINTA Peringkat 2 Pengaruh Oversampling pada Klasifikasi Hipertensi dengan Algoritma Naïve Bayes, Decision Tree, dan Artificial Neural Network (ANN)," *Masa Berlaku Mulai*, vol. 1, no. 3, pp. 635–641, 2017.

[16] E. Kurniawan, "Analisa Data Rekam Medis Menggunakan Teknik Data Mining Association Rules Dengan Algoritma Clustering .pdf," pp. 1–7.

[17] R. Arafiyah, F. Hermin, I. R. Kartika, A. Alimuddin, and I. Saraswati,

UNIVERSITAS MERCU BUANA

"Classification of Dengue Haemorrhagic Fever (DHF) using SVM, naive bayes and random forest," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 434, no. 1, 2018, doi: 10.1088/1757-899X/434/1/012070.

[18] U. Pujianto, A. L. Setiawan, H. A. Rosyid, and A. M. M. Salah, "Comparison of Naïve Bayes Algorithm and Decision Tree C4.5 for Hospital Readmission Diabetes Patients using HbA1c Measurement," *Knowl. Eng. Data Sci.*, vol. 2, no. 2, p. 58, 2019, doi: 10.17977/um018v2i22019p58-71.

[19] T. P. Ulfa, H Maria. Octaria, H. Sari, "Analisis Ketepatan Kode Diagnosa Penyakit Antara Rumah Sakit Dan BPJS Menggunakan ICD-10 Untuk Penagihan Klaim di Rumah Sakit Kelas C Sekota Pekan Baru Tahun 2016," *J. INOHIM*, vol. 5, p. 119, 2016, [Online]. Available: https://inohim.esaunggul.ac.id/index.php/INO/article/view/137.

[20] L. Masyfufah and I. Artikel, "Clustering Guna Menunjang Program Jaminan Balong Bendo Sidoarjo Clustering Bpjs ' S Inpatients By Using K -Means Clustering Method To Support the National Health Insurance Program At Anwar Medika Hospital Balong Bendo Sidoarjo," pp. 8–22, 2021.

[21] S. Informasi, U. P. Batam, K. Kunci, K. Kerja, and D. Mining, "Analisa dan Penerapan Algoritma C4 . 5 Dalam Data Mining Untuk Mengidentifikasi Faktor-Faktor Penyebab Kecelakaan Kerja Konstruksi PT . Arupadhatu Adisesanti," vol. 2, no. 1, pp. 36–41, 2017.

[22] M. T. Scholar and A. Chauhan, "Detection of Lung Cancer using Machine Learning Techniques Based on Routine Blood Indices," pp. 1–6, 2020.

[23] Angelia Fransiska, "METODE DATA MINING CLASSIFICATION." https://sis.binus.ac.id/2021/10/22/metode-data-mining-classification/.

UNIVERSITAS MERCU BUANA

KERTAS KERJA

Ringkasan

Kertas kerja ini merupakan material kelengkapan artikel jurnal dengan judul "Perbandingan Algoritma Naive Bayes dan Decision Tree untuk Klasifikasi Data Rekam Medis Berdasarkan Kode ICD-10 Di Rumah Sakit X Tangerang Selatan". Kertas kerja ini berisi semua material hasil penelitian Tugas Akhir. Di dalam kertas kerja ini disajikan beberapa bagian yang terdiri dari literature review, dataset yang digunakan, tahapan eksperimen, dan hasil eksperimen secara keseluruhan.

Bagian I membahas mengenai literature review yang berisi artikel jurnal yang menjadi dasar atau landasan dalam penelitian ini.Bagian II menjelaskan mengenai source code yang digunakan pada penelitian ini. Bagian III menjelaskan mengenai dataset yang digunakan, meliputi penjelasan dan sumber data. Bagian IV memuat tahapan eksperimen yang disajikan dalam gambar dengan penjelasan dari setiap tahapan. Bagian V merupakan bagian terakhir dari kertas kerja ini yang menjelaskan hasil keseluruhan dari eksperimen yang telah dilakukan, meliputi penjelasannya.

UNIVERSITAS
MERCU BUANA

https://lib.mercubuana.ac.id/