



**KLASIFIKASI *RANDOM FOREST* DENGAN PERBAIKAN *MISSING VALUE*
K-NEAREST NEIGHBOR IMPUTATION (KNNI) PADA PENYAKIT
DIABETES**

Skripsi

**Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Strata-1 Ilmu Komputer**

**Oleh
SARTIKA DEWI
1711016120019**

**PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU**

MEI 2023



**KLASIFIKASI *RANDOM FOREST* DENGAN PERBAIKAN *MISSING VALUE*
K-NEAREST NEIGHBOR IMPUTATION (KNNI) PADA PENYAKIT
DIABETES**

Skripsi

**Untuk Memenuhi Persyaratan
Dalam Menyelesaikan Strata-1 Ilmu Komputer**

**Oleh
SARTIKA DEWI
1711016120019**

**PROGRAM STUDI S-1 ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMBUNG MANGKURAT
BANJARBARU
MEI 2023**

SKRIPSI

KLASIFIKASI *RANDOM FOREST* DENGAN PERBAIKAN *MISSING VALUE K-NEAREST NEIGHBOR IMPUTATION (KNNI)* PADA PENYEKIT DIABETES

Oleh:

SARTIKA DEWI

NIM 1711016120019

Telah dipertahankan di depan Dosen Penguji pada Tanggal 23 Mei 2023

Susunan Dosen Penguji:

Pembimbing I



Triando Hamonangan Saragih,
S.Kom., M.Kom
NIP. 199308242019031012

Dosen Penguji I



Dodon Turianto Nugrahadi,
S.Kom., M.Eng
NIP. 198001122009121002

Pembimbing II



Radityo Adi Nugroho, S.T., M.Kom
NIP. 198212042008011006

Dosen Penguji II



Dwi Kartini, S.Kom., M.Kom.
NIP. 198704212012122003



PERNYATAAN

Dengan ini saya menyatakan bahwa dalam skripsi ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar keserjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Banjarbaru, 23 Mei 2023

Yang Menyatakan,



Sartika Dewi

NIM. 1711016120019

ABSTRAK

KLASIFIKASI *RANDOM FOREST* DENGAN PERBAIKAN *MISSING VALUE K-NEAREST NEIGHBOR IMPUTATION (KNNI)* PADA PENYAKIT DIABETES

(Oleh: Sartika Dewi; Pembimbing: Triando Hamonangan Saragih, S.Kom., M.Kom dan Radtyo Adi Nugroho, S.T., M.Kom; 2022; 67 halaman)

Pada saat melakukan survei data penghilangan nilai atau tidak lengkapnya data dapat terjadi, sehingga metode analisis yang tersedia hanya dapat bekerja dengan data lengkap. Pada penelitian ini melakukan penghilangan nilai dengan jumlah proporsi penghilangan yaitu 10%, 20% dan 30%. Tujuan dari penelitian ini yaitu untuk mengetahui hasil akurasi dari imputasi data hilang dengan K-NNI menggunakan klasifikasi *Random Forest* dan klasifikasi *Random Forest* tanpa imputasi. Metode penelitian dilakukan dengan Dataset publik *Pima Indian Diabetes* lalu dengan data yang lengkap dilakukan pengosongan data menggunakan MCAR (*Missing Completely At Random*). Kemudian, dilakukan 2 model yaitu yang pertama imputasi data hilang dengan K-NNI dan klasifikasikan dengan *Random Forest*. Kedua, klasifikasi *Random Forest* tanpa imputasi. Berdasarkan dari hasil akurasi K-NNI menggunakan klasifikasi *Random Forest* didapatkan akurasi terbaik pada penghilangan nilai sebanyak 10% dan nilai $K = 6$ yaitu sebesar 82,17% dengan presisi 76,71%, *recall* 70,00% dan *f1-Score* 73,20%. Kemudian untuk hasil akurasi dari klasifikasi *Random Forest* tanpa imputasi didapatkan pada penghilangan nilai 10% sebesar 80,00% dengan presisi 64,52%, *recall* 68,97% dan *f1-Score* 66,67%. Namun, pada pengujian missing data tanpa imputasi K-NNI menggunakan klasifikasi *Random Forest* memperoleh hasil yang tidak lebih baik dari hasil akurasi K-NNI menggunakan klasifikasi *Random Forest*.

Kata Kunci: *Pima Indian Diabetes, missing value, K-Nearest Neighbor Imputation, Random Forest.*

ABSTRACT

CLASSIFICATION OF RANDOM FORESTS WITH MISSING VALUE IMPUTATION OF K-NEAREST NEIGHBOR (KNNI) IN DIABETES DIABETES

(By: Sartika Dewi; Supervisor: Triando Hamonangan Saragih, S.Kom., M.Kom and Radtyo Adi Nugroho, S.T., M.Kom; 2022; 67 pages)

When conducting a data survey, omissions or incomplete data may occur, so that the available analytical methods can only work with complete data. In this study, devaluation was carried out with the proportion of omissions namely 10%, 20% and 30%. The purpose of this research is to find out the results of the accuracy of imputation of missing data with K-NNI using Random Forest classification and Random Forest classification without imputation. The research method was carried out using the Pima Indian Diabetes public dataset, and then with complete data, blank data was carried out using MCAR (Missing Completely At Random). Then, 2 models were carried out, namely the first was imputed missing data with K-NNI and classified with Random Forest. Second, the classification of Random Forest without imputation. Based on the results of the K-NNI accuracy using the Random Forest classification, the best accuracy was obtained for 10% omission and a $K = 6$ value of 82.17% with a precision of 76.71%, a recall of 70.00% and an f1-Score of 73.20 %. Then for the results of the accuracy of the Random Forest classification without imputation obtained at 10% omission of 80.00% with a precision of 64.52%, a recall of 68.97% and an f1-Score of 66.67%. However, the missing data test without K-NNI imputation using the Random Forest classification obtained results that were no better than the K-NNI accuracy results using the Random Forest classification.

Keywords: *Pima Indian Diabetes, missing value, K-Nearest Neighbor Imputation, Random Forest.*

PRAKATA

Puji syukur penulis panjatkan ke hadirat Allah SWT karena atas berkat rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi yang berjudul “**KLASIFIKASI RANDOM FOREST DENGAN PERBAIKAN MISSING VALUE K-NEAREST NEIGHBOR IMPUTATION (KNNI) PADA PENYAKIT DIABETES** ” untuk memenuhi syarat dalam menyelesaikan pendidikan program S1 Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Lambung Mangkurat. Tak lupa pula penulis panjatkan shalawat dan salam ke hadirat Rasulullah Muhammad SAW beserta para sahabat, keluarga, dan pengikut beliau hingga *yaumul qiyamah*.

Pada lembar ini penulis ingin menyampaikan ucapan terimakasih kepada pihak-pihak yang sangat mendukung penulis dalam pembuatan dan penyusunan skripsi ini, adapun yang dimaksud adalah sebagai berikut :

1. Keluarga terutama kepada Orang tua yang selalu memberikan bantuan, semangat, doa dan dukungan dalam proses penyelesaian skripsi ini.
2. Bapak Triando Hamonangan Saragih, S.Kom,. M.Kom selaku dosen pembimbing utama yang turut serta membantu dan meluangkan waktu demi kelancaran penyelesaian skripsi ini.
3. Bapak Radityo Adi Nugroho, S.T., M.Kom selaku dosen pembimbing pendamping yang turut serta membantu dan meluangkan waktu demi kelancaran dalam penyelesaian skripsi ini.
4. Bapak Irwan Budiman, S,T., M.Kom selaku Koordinator Program Studi Ilmu Komputer FMIPA ULM, atas bantuan dan izin beliau skripsi ini dapat diselesaikan.
5. Seluruh Dosen dan staf Program Studi Ilmu Komputer FMIPA UNLAM atas ilmu dan bantuan yang diberikan selama ini yang sangat bermanfaat.
6. Teman-teman keluarga Ilmu Komputer angkatan 2017, terima kasih untuk canda, tawa, perjuangan yang sudah dilewati bersama, untuk semua kenangan manis yang telah terukir selama ini. Senang bisa menjadi salah satu bagian dari kehidupan kalian.

7. Terima kasih juga kepada teman-teman yang sudah mendukung saya dalam menyelesaikan penelitian ini, terima kasih kepada Noor Hidayah yang sudah banyak membantu dalam segala hal, terima kasih juga kepada Syaoki Paradisa yang sudah membantu saya dalam proses penelitian saya, terima kasih banyak juga kepada Sulastri Nor Indah Sari, Mia Lidiyani, Regina, dan Nurhuda dan Muhammad Nor Azmi atas bantuan dan dukungannya selama ini.
8. Semua pihak yang tidak dapat disebutkan satu persatu yang telah turut membantu dalam penyelesaian skripsi ini.

Akhir kata penulis menyadari sepenuhnya bahwa penulisan ini jauh dari sempurna, namun penulis mengharapkan bantuan berupa saran dan kritik yang membangun dari semua pihak demi kesempurnaan dan mutu penulisan skripsi ini.

Semoga tulisan ini dapat bermanfaat bagi ilmu pengetahuan dan pembaca khususnya serta mendapat keridhaan Allah SWT.

Banjarbaru, 23 Mei 2023

Penulis

DAFTAR ISI

| | |
|---|-------------|
| HALAMAN JUDUL | i |
| LEMBAR PENGESAHAN | ii |
| PERNYATAAN..... | iii |
| ABSTRAK | iv |
| ABSTRACT..... | v |
| PRAKATA..... | vi |
| DAFTAR ISI..... | viii |
| DAFTAR TABEL | x |
| DAFTAR GAMBAR..... | xii |
| DAFTAR LAMPIRAN..... | xiv |
| BAB I PENDAHULUAN..... | 1 |
| 1.1 Latar Belakang..... | 1 |
| 1.2 Rumusan Masalah..... | 2 |
| 1.3 Batasan Masalah | 3 |
| 1.4 Tujuan | 3 |
| 1.5 Manfaat Penelitian..... | 3 |
| BAB II TINJAUAN PUSTAKA..... | 4 |
| 2.1 Kajian Terdahulu | 4 |
| 2.2 Landasan Teori | 6 |
| 2.2.1 Diabetes | 6 |
| 2.2.2 Machine Learning | 7 |
| 2.2.3 Data Mining | 8 |
| 2.2.4 Data Preprocessing | 10 |
| 2.2.5 Missing Value | 10 |
| 2.2.6 K-Nearest Neighbor Imputation (KNNI)..... | 12 |
| 2.2.7 Klasifikasi Random Forest..... | 14 |
| 2.2.8 Split Validation..... | 18 |
| 2.2.9 Performa Klasifikasi | 18 |

| | |
|---|-----------|
| 2.2.10 Pemrograman R | 21 |
| BAB III METODE PENELITIAN | 24 |
| 3.1 Alat Penelitian | 24 |
| 3.2 Bahan Penelitian | 24 |
| 3.3 Variabel Penelitian..... | 24 |
| 3.4 Prosedur Penelitian | 25 |
| BAB IV HASIL DAN PEMBAHASAN | 27 |
| 4.1 Hasil..... | 27 |
| 4.1.1 Pengumpulan Data..... | 27 |
| 4.1.2 Membangkitkan Missing Data Secara Acak..... | 28 |
| 4.1.3 Penanganan Missing Value Dengan K-Nearest Neighbor Imputation | 30 |
| 4.1.4 Klasifikasi Random Forest..... | 32 |
| 4.1.5 Hasil Evaluasi | 44 |
| 4.2 Pembahasan | 59 |
| BAB V PENUTUP..... | 71 |
| 5.1 Kesimpulan..... | 71 |
| 5.2 Saran | 72 |
| DAFTAR PUSTAKA | 73 |
| LAMPIRAN | |

DAFTAR TABEL

| Tabel | Halaman |
|--|----------------|
| Tabel 1. Keaslian penelitian..... | 5 |
| Tabel 2. Perancangan penelitian | 5 |
| Tabel 3. Contoh Data Kosong..... | 12 |
| Tabel 4. <i>Confusion matrix multiclass</i> | 19 |
| Tabel 5. Deskripsi Dataset | 27 |
| Tabel 6. Sampel Dataset..... | 28 |
| Tabel 7. Hasil Penghilangan Data Berdasarkan Proporsi 10% | 29 |
| Tabel 8. Hasil Penghilangan Data Berdasarkan Proporsi 20% | 29 |
| Tabel 9. Hasil Penghilangan Data Berdasarkan Proporsi 30% | 30 |
| Tabel 10. Data untuk Perhitungan Model K-Nearest Neighbor Imputation | 30 |
| Tabel 11. Contoh Hasil Pengujian Nilai K = 2 Dengan Persentase Penghilangan Nilai 10% | 32 |
| Tabel 12. Data untuk model Random Forest | 34 |
| Tabel 13. Nilai entropy dan gain pada pohon pertama | 35 |
| Tabel 14. Data dari atribut BloodPressure yang bernilai ≥ 43 | 36 |
| Tabel 15. Nilai entropy dan gain pada node 1.2 | 36 |
| Tabel 16. Data dari atribut Pregnancies yang bernilai ≥ 7 | 37 |
| Tabel 17. Nilai entropy dan gain pada node 2.2 | 38 |
| Tabel 18. Data ke dua untuk model Random Forest..... | 39 |
| Tabel 19. Data ke tiga model Random Forest..... | 41 |
| Tabel 20. Hasil majority voting | 43 |
| Tabel 21. Confusion matrix Random Forest..... | 44 |
| Tabel 22. Hasil Confusion Matrix dengan nilai K berdasarkan proporsi 10% | 44 |
| Tabel 23. Hasil Confusion Matrix dengan nilai K berdasarkan proporsi 20% | 47 |
| Tabel 24. Hasil Confusion Matrix dengan nilai K Berdasarkan proporsi 30% | 50 |
| Tabel 25. Hasil Dari Klasifikasi Random Forest Dengan Hasil Pengujian Nilai K Perbaikan Missing Imputasi Menggunakan KNNI..... | 52 |

| | |
|--|----|
| Tabel 26. Hasil dari klasifikasi Random Forest tanpa perbaikan missing imputasi menggunakan KNNI | 53 |
| Tabel 27. Performa AUC klasifikasi <i>Random Forest</i> tanpa perbaikan <i>missing</i> data menggunakan KNNI | 58 |
| Tabel 28. Performa AUC klafikasi <i>Random Forest</i> dengan perbaikan <i>missing</i> data menggunakan KNNI | 58 |
| Tabel 29 Performa AUC klafikasi <i>Random Forest</i> dengan perbaikan <i>missing</i> data menggunakan KNNI | 62 |
| Tabel 30. Confusion matrix klasifikasi random forest tanpa perbaikan imputasi dengan penghilangan 10% | 66 |
| Tabel 31. Confusion matrix klasifikasi random forest tanpa imputasi dengan penghilangan nilai 20% | 67 |
| Tabel 32. Confusion matrix klasifikasi random forest tanpa imputasi dengan penghilangan nilai 30% | 68 |
| Tabel 33. Performa AUC klasifikasi <i>Random Forest</i> tanpa perbaikan <i>missing</i> data menggunakan KNNI | 68 |

DAFTAR GAMBAR

| Gambar | Halaman |
|--|----------------|
| Gambar 1 Alur <i>Machine Learning</i> | 7 |
| Gambar 2 Klasifikasi <i>Random Forest</i> | 17 |
| Gambar 3 Ilustrasi <i>Split Validation</i> | 18 |
| Gambar 4. Alur Penelitian..... | 25 |
| Gambar 5 Root node pohon keputusan <i>Random Forest</i> | 35 |
| Gambar 6 Node 1.2 pohon keputusan <i>Random Forest</i> | 37 |
| Gambar 7 Pohon keputusan pertama <i>Random Forest</i> | 38 |
| Gambar 8 Pohon keputusan kedua <i>Random Forest</i> | 40 |
| Gambar 9 Pohon keputusan ketiga <i>Random Forest</i> | 41 |
| Gambar 10 Evaluasi model dari nilai K berdasarkan 10% | 47 |
| Gambar 11 Evaluasi model dari nilai K berdasarkan 20% | 49 |
| Gambar 12 Evaluasi model dari nilai K berdasarkan 30% | 52 |
| Gambar 13 Grafik hasil evaluasi klasifikasi <i>random forest</i> dengan perbaikan KNNI | 53 |
| Gambar 14 Hasil evaluasi <i>random forest</i> tanpa <i>missing</i> imputasi menggunakan KNNI | 57 |
| Gambar 15 Hasil evaluasi klasifikasi <i>random forest</i> dengan perbaikan KNNI | 62 |
| Gambar 16 Hasil performa AUC klasifikasi <i>Random Fores</i> dengan perbaikan <i>missing</i> data menggunakan KNNI berdasarkan persentase penghilangan 10% | 63 |
| Gambar 17 Hasil performa AUC klasifikasi <i>Random Fores</i> dengan perbaikan <i>missing</i> data menggunakan KNNI berdasarkan persentase penghilangan 20% | 64 |
| Gambar 18 Hasil performa AUC klasifikasi <i>Random Fores</i> dengan perbaikan <i>missing</i> data menggunakan KNNI berdasarkan persentase penghilangan 30%..... | 64 |
| Gambar 19 Hasil evaluasi klasifikasi <i>random forest</i> tanpa imputasi | 66 |
| Gambar 20 Hasil performa AUC klasifikasi <i>Random Fores</i> tanpa perbaikan <i>missing</i> data menggunakan KNNI berdasarkan persentase penghilangan 10% | 69 |
| Gambar 21 Hasil performa AUC klasifikasi <i>Random Fores</i> tanpa perbaikan <i>missing</i> data menggunakan KNNI berdasarkan persentase penghilangan 20% | 69 |

Gambar 22 Hasil performa AUC klasifikasi *Random Fores* tanpa perbaikan *missing* data menggunakan KNNI berdasarkan persentase penghilangan 30% 70

DAFTAR LAMPIRAN

Lampiran

Lampiran 1. Tabel Performa Akhir Klasifikasi Random Forest Dengan Menggunakan K-Nearest Neighbor Imputation Dengan Proporsi 10%

Lampiran 2. Tabel Performa Akhir Klasifikasi Random Forest Dengan Menggunakan K-Nearest Neighbor Imputation Dengan Proporsi 20%

Lampiran 3. Tabel Performa Akhir Klasifikasi Random Forest Dengan Menggunakan K-Nearest Neighbor Imputation Dengan Proporsi 30%

Lampiran 4. Tabel Performa Akhir Klasifikasi Random Forest tanpa perbaikan missing imputasi Menggunakan K-Nearest Neighbor Imputation Dengan Proporsi 10%, 20%, dan 30%

Lampiran 5. Source Code klasifikasi random forest dengan perbaikan missing imputasi menggunakan K-Nearest Neighbor Imputation

Lampiran 6. Source Code klasifikasi random forest tanpa perbaikan missing imputasi menggunakan K-Nearest Neighbor