PROJEK AKHIR UAS

BIG DATA AND DATA MINING (ST168)

Deteksi Penyakit Parkinson Berdasarkan Pengukuran Suara Biomedis Menggunakan Algoritma KNN

Disusun oleh

22.11.4966

Waode Nurdinia Anisa

IF 07

PROGRAM STUDI S1 INFORMATIKA

FAKULTAS ILMU KOMPUTER

UNIVERSITAS AMIKOM YOGYAKARTA

2025

1. PENDAHULUAN

Penyakit parkinson adalah gangguan yang terdapat pada sisten saraf pusat yang dapat mempengaruhi sisem motorik, diagnosa yang di temukan pada gejala penyakit ini serupa dengan penyakit lain sehingga cukup sulit untuk di prediksi [1]. Beberapa gejala yang terjadi pada penderita penyakit ini di antaranya pada gangguan intelek dan tingkah laku, demensia, penurunan daya ingat, kelemahan otot, katalepsi, dan tremor [2]. Penyakit ini kebanyakan terdominasi pada laki-laki di bandingkan wanita dengan perbandingan 3 banding 2 dan perkiraan muncul pada usia 40-70 tahun, rata-ratan lansia diatas 60 tahun . di perkirakan penderita penyakit parkinson di indonesia sendir terdapat 200.000-400.000 [3]

Organisasi Kesehatan Dunia telah memprediksi terdapat satu juta orang meniggal dunia karena peyakit parkinsom setiap tahun dan jumlah penderita penyakit ini terjadi penikatan hingg empat kali lipat dari tahun-tahun sebelumnya [4]. Untuk mengurangi keparahan pada gejala penyakit ini, dengan menerapkan proses deteksi dini dimana untuk mengidentifikasi penyakit tahap awal untuk mengatasi kemunculan penyakit menjadi lebih parah. Dengan melakukan pengobatan yang lebih cepat dan efektif untuk memperlambat penyakit berkembang [5].

Pada penelitian ini, deteksi penyakit parkinson menerapkan *Algoritma K-Nearest Neighbor* (KNN). *Algoritma K-Nearest Neighbor* (KNN) merupakan salah satu algoritma yang cukup sederhana dangan mengklasifikasi suatu objek tetangga terdekat dan lebih mayoritas [6]. Algoritma KNN telah banyak di gunakan pada banyak masalah klasifikasi sehingga menjadi algoritma yang lebih banyak diteliti [7]. Penerapan pada algoritma KNN pada dataset ini menghasilkan akurasi yang cuku baik yaitu 99%, dengan kelas 1 memiliki Recall sempurna (1.00) yang berarti model tidak melewati satupun sampel dari kelas 1, sedangkan kelas 0 sedikit kurang baik dalam Recall (98%), tetapi Precision-nya tetap tinggi (99%). Pada nilai F1-Score mencapai 98% dan 99%, sehingga model ini sangat seimbang antara Precision dan Recal, menujukkan kinerja yang stabul dan konsisten [8].

2. PROFILE DATASET

pada penelitian ini dataset yang digunakan merupakan dataset Kumpulan Data Penyakit Parkinson yang dilakukan dengan berbagai pengukuran suara biomedis. Tujuan utama dari data ini yang sehat dan pengidap penyakit parkinson. Dataset ini memiliki 1727 baris data dan 24 fitur terkait penyakit parkinson. Dataset ini di ambil dari kaggle dan di unggah oleh DR. James Bond dan telah di perbaharui 9 bulan yang lalu.

Fitur yang terdapat pada dataset dan deskripsi :

- nama - nama subjek ASCII dan nomor rekaman
- MDVP:Fo(Hz) - Frekuensi dasar vokal rata-rata
- MDVP:Fhi(Hz) - Frekuensi dasar vokal maksimum
- MDVP:Flo(Hz) - Frekuensi dasar vokal minimum
- MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP - Beberapa ukuran variasi dalam frekuensi dasar
- MDVP:Shimmer,MDVP:Shimmer(dB),Shimmer:APQ3,Shimmer:APQ5,MDVP:APQ,Shimmer:DDA - Beberapa ukuran variasi dalam amplitudo
- NHR, HNR - Dua ukuran rasio komponen noise terhadap tonal dalam

- status suara - Status kesehatan subjek (satu) - Parkinson, (nol) - sehat
- RPDE, D2 - Dua ukuran kompleksitas dinamis nonlinier
- DFA - Eksponen skala fraktal sinyal
- spread1, spread2, PPE - Tiga ukuran variasi frekuensi dasar nonlinier

3. DATA PREPROCESSING

Proses Preprocessing data dengan menggunakan fitur selection dan normalization dengan alas an sebagai berikut :

- **Feature Selection**: Langkah ini bertujuan untuk memilih fitur-fitur yang memiliki hubungan kuat dengan target, sehingga dapat meningkatkan akurasi model sekaligus mengurangi redundansi data.
- **Normalization**: Langkah ini dilakukan untuk menyamakan rentang nilai antar fitur yang bervariasi secara signifikan. Hal ini memastikan bahwa semua fitur memiliki skala yang seragam, sehingga model dapat beroperasi secara lebih efektif.
- **Handling Imbalanced Data :** proses menangani ketidakseimbangan pada dataset, di mana jumlah sampel dari satu kelas jauh lebih besar atau lebih kecil dibandingkan kelas lainnya.

4. EXPLORATORY DATA ANALYSIS

Dalam analisis EDA, beberapa aspek penting yang diperiksa meliputi: mengevaluasi keberadaan missing values pada setiap fitur, memverifikasi tipe data dari masing-masing fitur, meninjau tampilan data secara keseluruhan, menganalisis bentuk data (jumlah baris dan kolom), serta mengidentifikasi korelasi antar fitur.

a. Terdapat data yang terdi dari 1727 dan 24 fitur

```
[ ]  print(df.shape)

(1727, 24)
```

b. Pada dataset ini berkaitan dengan pengukuran suara, seperti pada analisi kesehatan suara atau diagnosis penyakit. Terdapat tipe numerik (float dan integer) dengan mayoritas kolom pada nila desimal (float64), sedangakan pada kolom status terdapat data kategori numerik(integer) dan kolom name terdapat fungsi identifier dengan tipe string(object).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1727 entries, 0 to 1726
Data columns (total 24 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   name              1727 non-null   object
 1   MDVP:Fo(Hz)       1727 non-null   float64
 2   MDVP:Fhi(Hz)      1727 non-null   float64
 3   MDVP:Flo(Hz)      1727 non-null   float64
 4   MDVP:Jitter(%)    1727 non-null   float64
 5   MDVP:Jitter(Abs)  1727 non-null   float64
 6   MDVP:RAP          1727 non-null   float64
 7   MDVP:PPQ          1727 non-null   float64
 8   Jitter:DDP        1727 non-null   float64
 9   MDVP:Shimmer      1727 non-null   float64
 10  MDVP:Shimmer(dB)  1727 non-null   float64
 11  Shimmer:APQ3      1727 non-null   float64
 12  Shimmer:APQ5      1727 non-null   float64
 13  MDVP:APQ          1727 non-null   float64
 14  Shimmer:DDA       1727 non-null   float64
 15  NHR               1727 non-null   float64
 16  HNR               1727 non-null   float64
 17  status            1727 non-null   int64
 18  RPDE              1727 non-null   float64
 19  DFA               1727 non-null   float64
 20  spread1           1727 non-null   float64
 21  spread2           1727 non-null   float64
 22  D2                1727 non-null   float64
 23  PPE               1727 non-null   float64
dtypes: float64(22), int64(1), object(1)
memory usage: 323.9+ KB
```

c. Pada data yang di tampilkan tidak di temukan yang mengalami missing valeu

	0
name	0
MDVP:Fo(Hz)	0
MDVP:Fhi(Hz)	0
MDVP:Flo(Hz)	0
MDVP:Jitter(%)	0
MDVP:Jitter(Abs)	0
MDVP:RAP	0
MDVP:PPQ	0
Jitter:DDP	0
MDVP:Shimmer	0
MDVP:Shimmer(dB)	0
Shimmer:APQ3	0
Shimmer:APQ5	0
MDVP:APQ	0
Shimmer:DDA	0
NHR	0
HNR	0
status	0
RPDE	0
DFA	0
spread1	0
spread2	0
D2	0
PPE	0

dtype: int64

d. Pada dataset ini terjadi tidak ketidak seimbangan data pada kolom status dimana nilai pada label 1 lebih banyak dibandingkan label 0 dengan rasio data kira- 3:1. Sehingga pada penelitian langkah yang di lakukan dengan melakukan teknik SMOTE (Synthetic Minority Oversampling Technique) guna untuk dilakukan oversampling pada class 0 dan undersampling pada class 1.

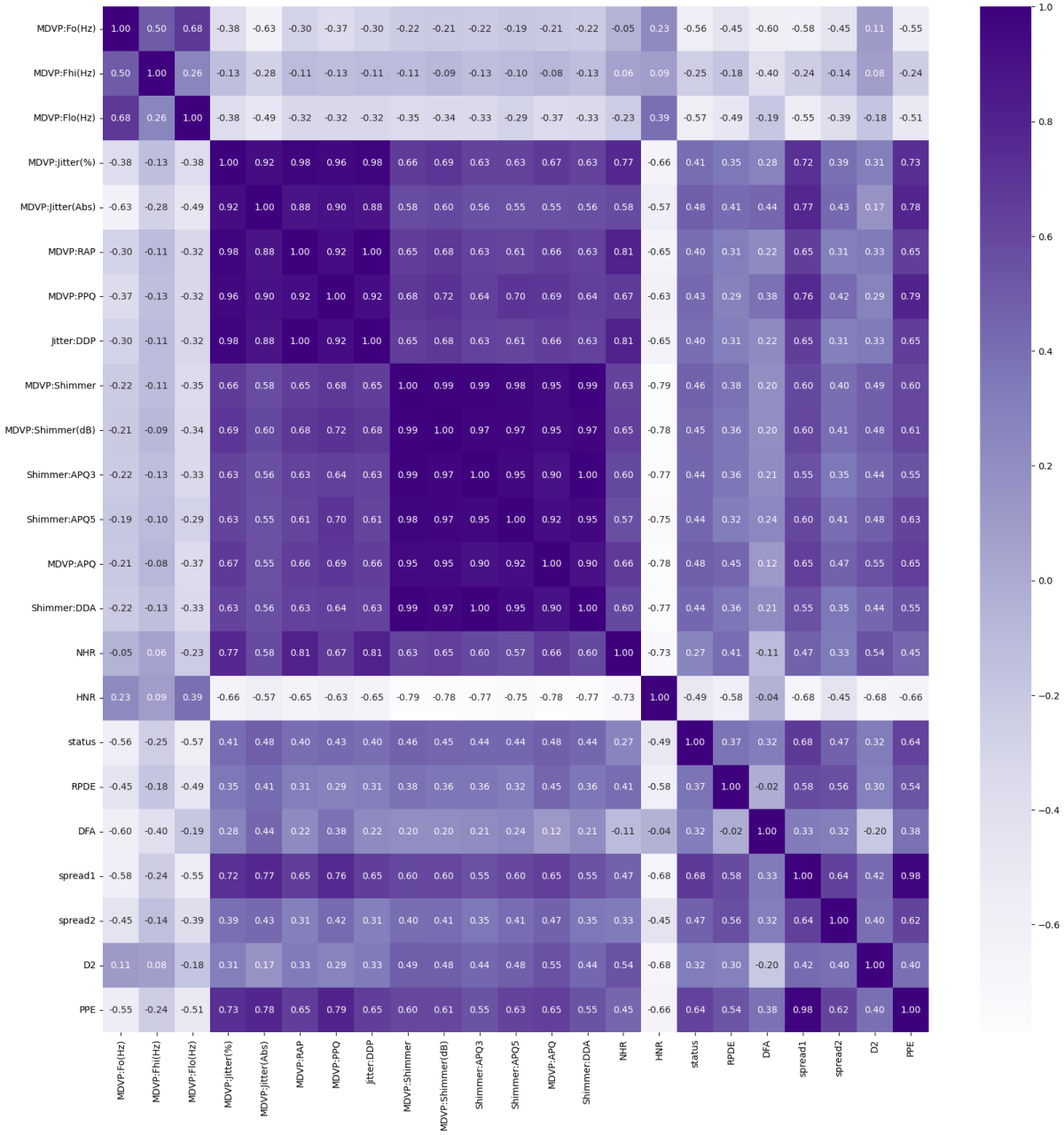
status	count
1	1289
0	438

dtype: int64

SMOTE membuat data set seimbang dengan mensistensi sampel baru pada class 0 yang berdasarkan tetangga terdekat untuk meningkatkan performa pada kelas mayoritas.

```
Distribusi sebelum SMOTE: Counter({1: 1289, 0: 438})
Distribusi setelah SMOTE: Counter({1: 1289, 0: 1289})
```

e. Tampilan pada fitur yang di gunakan sebelum melakukan proses seleksi fitur



5. SELEKSI FITUR

Dalam proses seleksi fitur, digunakan matriks korelasi untuk menilai relevansi fitur terhadap target variabe. Fitur dengan nilai korelasi dalam rentang -0.1 hingga 0.1 dianggap memiliki hubungan yang lemah atau tidak signifikan terhadap target. Oleh sebab itu, fitur-fitur tersebut dihapus karena kontribusinya terhadap model dianggap minimal. Tujuan seleksi fitur ini adalah untuk meningkatkan performa model dengan hanya mempertahankan fitur-fitur yang memiliki hubungan lebih kuat dengan target variabel.

```
MDVP:Fo(Hz)        -0.561985
MDVP:Fhi(Hz)       -0.248930
MDVP:Flo(Hz)       -0.574260
MDVP:Jitter(%)      0.410573
MDVP:Jitter(Abs)    0.483587
MDVP:RAP            0.402748
MDVP:PPQ            0.431241
Jitter:DDP          0.402837
MDVP:Shimmer        0.463516
MDVP:Shimmer(dB)    0.451633
Shimmer:APQ3        0.439272
Shimmer:APQ5        0.440638
MDVP:APQ            0.480911
Shimmer:DDA         0.439276
NHR                 0.270468
HNR                -0.491311
RPDE                0.374460
DFA                 0.318137
spread1             0.675079
spread2             0.469929
D2                  0.323753
PPE                 0.636712
dtype: float64
```

Pada korelasi ini menunjukkan hubungan linear antara dua variabel. Seperti pada MDVP:Fo(HZ) dengan hubungan linear negatif sedang dengan target dan spread1 denngan menunjukkan hubungan linear positif yang kuat dengan target

Pada seleksi fiturnya menunjukkan proses fitur berdasarkan nilai absolut korelasi yang signifikan dengan target variabel sehingga model lebih efektif. Dari hasil seleksi, fitur seperti **MDVP:Fo(Hz)**, **spread1**, dan **PPE** dianggap signifikan dan layak digunakan dalam pelatihan model.

6. MODELING

Dalam menganalisis data menggunkan metode *Algoritma K-Nearest Neighbor* (KNN), mennghasilkan akurasi 90%. Terdapat perbandingan 3:1 sebelum di Sebelum di SMOTE dan setelah di SMOTE, agar dataset menjadi seimbang dengan hasil perbanding 1:1.

a. Link model github : https://github.com/WaodeAnisaNurdinia/PreprocessingModelKNN
b. Link Google Colab : https://colab.research.google.com/drive/1MqWVVUJL0HC4m8tUkZt1SbDN4ls_T2qY?usp=sharing

7. EVALUASI MODEL

Dalam melakukan evaluasi dengan menggunakann model *Algoritma K-Nearest Neighbor* (KNN), model memiliki performa yang sangat baik dengan mennghasilkan akurasi tinggi (99%), Precision, recall, dan F1-score untuk kedua kelas juga sangat baik, menunjukkan keseimbangan performa model untuk keduanya.

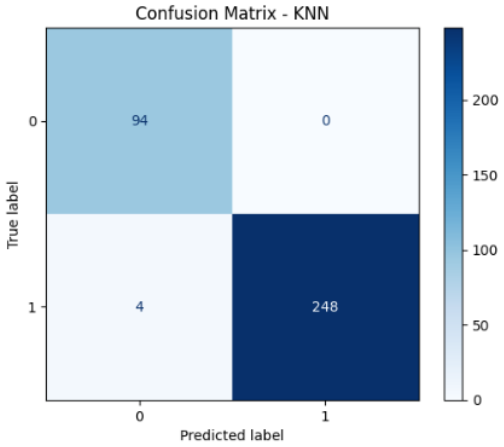
```
Confusion Matrix:
[[ 94   0]
 [  4 248]]

Classification Report:
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        94
           1       1.00      0.98      0.99       252

    accuracy                           0.99       346
   macro avg       0.98      0.99      0.99       346
weighted avg       0.99      0.99      0.99       346
```

Hanya ada 4 kasus salah prediksi (False Negatives), dan tidak ada False Positives, yang menunjukkan bahwa model hampir sempurna. Pada tampilan Cunfusiom Matrix menghasilkan jumlah prediksi yang benar 248 pada TP class 1 dan 94 pada TN class 0, sedangkan pada FP terdapat 0 dengan jumlah kasus 0 yang salah prediksi sebagi class 1, dan FN terdapat 4 dengan Jumlah kasus kelas 1 yang salah diprediksi sebagai kelas 0.



8. ANALISA DAN PEMBAHASAN

Setelah melakukan analisi dengan menggunakan model *Algoritma K-Nearest Neighbor* (KNN) pada dataset ini menunjukkan hasil yang cukup baik dengan nilai akurasi mencapai 99%, dengan pemilihan fitur dan proses preprocessing. Hanya saja pada dataset ini tidak seimbang antara data clas 0 dan class 1. Dengan menggunakan teknik SMOTE dengan dilakukan oversampling pada class 0 dan undersampling pada class 1, guna membuat data set seimbang dengan mensistensi sampel baru pada class 0 yang berdasarkan tetangga terdekat.

9. KESIMPULAN

algoritma K-Nearest Neighbor (KNN) menunjukkan performa yang sangat baik dalam mendeteksi penyakit Parkinson dengan akurasi 99%. Model ini stabil dan konsisten, meski dataset awal tidak seimbang. Penerapan teknik SMOTE berhasil menyeimbangkan data, meningkatkan kualitas prediksi, dan mendukung deteksi dini yang efektif untuk mencegah keparahan gejala.

10. Referensi

[1] Kurnia, D., Mazdadi, M. I., Kartini, D., Nugroho, R. A., & Abadi, F. (2023). Seleksi Fitur dengan Particle Swarm Optimization pada Klasifikasi Penyakit Parkinson Menggunakan XGBoost. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *10*(5), 1083-1094.

[2] Diantika, Dictia. "Implementasi Naïve Bayes Classifier untuk Penyakit Parkinson." PhD diss., Universitas Islam Negeri Sultan Syarif Kasim Riau, 2020.

[3] Sany, F. A. (2024). *Pengembangan Aplikasi Parkinfo untuk Lansia Penyandang Penyakit Parkinson* (Doctoral dissertation, Universitas Islam Indonesia).

[3] Desiani, A., Narti, N., Ramayanti, I., Arhami, M., & Irmeilyana, I. (2023). DIAGNOSA PENYAKIT PARKINSON DENGAN ALGORITMA K-NEAREST NEIGHTBOR DAN DECISION TREE C4. 5. *Jurnal Simantec*, *12*(1), 47-58.

[4] Aprilitaz, W., Akbar, R., & Prayogi, R. C. (2023, August). Komparasi Algoritma K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi Penyakit Parkinson: Comparison of K-Nearest Neighbor (KNN) and Naive Bayes Algorithms in the Classification of Parkinson's Disease. In *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat* (pp. 188-193).

[5] Wulandari, K. A., Nugraha, A., Luthfiarta, A., & Nisa, L. R. (2024). Peningkatan Akurasi Deteksi Dini Penyakit Parkinson melalui Pendekatan Ensemble Learning dan Seleksi Fitur Optimal. *Edumatic: Jurnal Pendidikan Informatika*, *8*(2), 575-584.

[6] Ridho, Wahyu Anwar, Triastuti Wuryandari, and Arief Rachman Hakim. "PERBANDINGAN KINERJA METODE KLASIFIKASI K-NEAREST NEIGHBOR DAN SUPPORT VECTOR MACHINES PADA DATASET PARKINSON." *Jurnal Gaussian* 12, no. 3 (2024): 372-381.

[7] Desiani, A., Narti, N., Ramayanti, I., Arhami, M., & Irmeilyana, I. (2023). DIAGNOSA PENYAKIT PARKINSON DENGAN ALGORITMA K-NEAREST NEIGHTBOR DAN DECISION TREE C4. 5. *Jurnal Simantec*, *12*(1), 47-58.

[8] MAHIDA, ALGA. "PENERAPAN REDUKSI DIMENSI DENGAN LINEAR DISCRIMINANT ANALYSIS PADA KLASIFIKASI PENYAKIT ARITMIA."

Kutipan Dataset : Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Kesesuaian pengukuran disfonia untuk telemonitoring penyakit Parkinson's', IEEE Transactions on Biomedical Engineering (akan muncul).