Journal of Smart and Intelligent Technology (SIT)
Vol.1, No.1 (2024)

OPTIMIZATION OF PARAMETERS K IN THE K-NEAREST NEIGHBOUR ALGORITHM FOR CLASSIFICATION OF DIABETES DISEASE BASED ON PYTHON

Grasberg Nahumarury [1)] , Anas Nasrullah [2)]

[1)] *Informatika Institut Teknologi Tangerang Selatan*
[2)] *Sistem Informasi Institut Teknologi Tangerang Selatan*
email : nahumarurygrasberg@gmail.com[1)], annas@itts.ac.id[2)]

Abstract

Diabetes doesn't just cause premature death worldwide. This disease is also a major cause of blindness, heart disease, and kidney failure. The International Diabetes Federation (IDF) organization estimates that at least 463 million people aged 20-79 years in the world have diabetes in 2019, or the equivalent of a prevalence rate of 9.3% of the total population at the same age. The research objective is to optimize the k parameter in the k-NN algorithm for python-based diabetes classification. This research was conducted using the experimental method. This experimental method was carried out by researchers by changing the k parameter with a value of 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, and 49 *and getting the research results for optimizing the value of k = 11 having the highest accuracy of 0.9617.*

Keywords **:**
Optimization, Diabetes, Classification, k-NN

1. Introduction

Diabetes not only causes premature deaths throughout the world. This disease is also a major cause of blindness, heart disease and kidney failure. The International Diabetes Federation (IDF) organization estimates that at least 463 million people aged 20-79 years in the world suffered from diabetes in 2019, or the equivalent of a prevalence rate of 9.3% of the total population of the same age. Based on gender, IDF estimates that the prevalence of diabetes in 2019 is 9% in women and 9.65% in men. The prevalence of diabetes is estimated to increase as the population increases to 19.9% or 111.2 million people aged 65-79 years. The figure is predicted to increase to reach 578 million in 2030 and 700 million in 2045.

The research produced the best K in the K=13 experiment with an accuracy of 75.14%. K=13 is the most optimal k value among KNN classification experiments using values K=1 to K=49 using rapid miner [1]. This research presents two classification methods, namely the Backpropagation method and Learning Vector Quantization for the problem of classifying diabetes mellitus. The conclusion was that in the classification of diabetes mellitus the Backprogation method provides better performance than LVQ [2]. By using a 60/40 data split, the Naïve Bayes algorithm produces an accuracy of 0.7608. Meanwhile, the results of Naïve Bayes which were boosted using the Adaboost algorithm were 0.7694 using Python [3]. The C4.5 algorithm is included in the classification algorithm which produces decision trees and can be processed with discrete and numerical data, besides that the C4.5 algorithm can produce a way that is easy to interpret in research. The accuracy of the C4.5 algorithm is 74.08% using Python and Rapid. miner[4].

To optimize the k parameters in the k-NN algorithm for diabetes classification, the researchers here used a dataset sourced from www.kaggle.com which amounted to 100,000 data and carried out tests using k parameters with values of 3, 5, 7, 9, and 11 using the application python and get research results for optimizing the value k=11 which has the highest accuracy of 0.9617.

2. Literature review

2.1. Studi Overview

- The research produced the best K in the K=13 experiment with an accuracy of 75.14%. K=13 is the most optimal k value among KNN classification experiments using values K=1 to K=49 using rapid miner [1].
- This research presents two classification

methods, namely the Backpropagation method and Learning Vector Quantization for the problem of classifying diabetes mellitus. The conclusion was that in the classification of diabetes mellitus the Backprogation method provides better performance than LVQ [2].

- By using a 60/40 data split, the Naïve Bayes algorithm produces an accuracy of 0.7608. Meanwhile, the results of Naïve Bayes which were boosted using the Adaboost algorithm were 0.7694 using Python. [3].
- The C4.5 algorithm is included in the classification algorithm which produces decision trees and can be processed with discrete and numerical data, besides that the C4.5 algorithm can produce a way that is easy to interpret in research. The accuracy of the C4.5 algorithm is 74.08% using Python and Rapid miners [4].

2.2. K-Nearest Neighbor (k-NN)

The KNN algorithm calculation uses the distance between all training data and testing data, and then the closest data or data that has the most similarities will be taken for classification. The KNN algorithm calculation uses the distance between all training data and testing data, and then the closest data or data that has the most similarities will be taken for classification [6].

2.3 Python

- Python is a programming language that allows you to work faster and integrate your systems more effectively.

How to Install Python:

- Open the www.python.org website page in the browser.
- Then select the download menu > windows
- Then select the available Python version then download the "Windows Python Installer" file.
- Then wait until the download process is complete.
- Then install the python file that has been successfully downloaded.
- Then during the installation process, make sure to check the "add python ... to path ...." to be operated via command prompt.
- Then wait until the installation process is complete.

2.4. Jupyter Notebook

Jupyter Notebook is a native web application for creating and sharing computational documents. It offers a simple, efficient, and document-centric experience.

How to Install Jupiter Notebook:

- Click the start button then open the command prompt
- Then type "pip install notebook" in the command prompt then press enter.
- Then wait until the installation process is complete.

2.5. NumPy

NumPy is a basic package for scientific computing in Python. It is a Python library that provides multidimensional array objects, various derived objects (such as arrays and masked matrices), and various routines for fast operations on arrays, including mathematics, logic, shape manipulation, sorting, selection, I/O, Fourier transforms discrete, basic linear algebra, basic statistical operations, random simulation, and more.

How to Install Numpy:

- Click the start button then open the command prompt
- Then type "pip install numpy" in the command prompt then press enter.
- Then wait until the installation process is complete.

Pandas is a fast, powerful, flexible, and easy-to-use open source data analysis and manipulation tool, built on the Python programming language.

How to Install Pandas:

- Click the start button then open the command prompt
- Then type "pip install pandas" in the command prompt then press enter.
- Then wait until the installation process is complete.

2.6. Scikit-learn

Scikit-learn is an open source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection, model evaluation, and many other utilities.

- Simple and efficient tool for predictive data analysis.
- It is accessible to everyone, and can be reused in various contexts.
- Built on top of NumPy, SciPy, and matplotlib.

- Open Source, can be used commercially - BSD license.
- How to Install Matplotlib on Windows:
- Click the start button then open the command prompt
- Then type “pip install scikit-learn” in the command prompt then press enter.
- Then wait until the installation process is complete.

3. Research methods

The purpose of this research is to optimize the k parameter in the k-Nearest Neighbor algorithm for Python-based diabetes classification to make it more accurate. To find out, we must test the k parameter in the k-Nearest Neighbor algorithm. This is done so that we can find out the level of accuracy of each k parameter.

This research was conducted using experimental methods. This experimental method was carried out by researchers by testing the parameter k with values 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41 , 43, 45, 47, and 49 according to the needs of the problems faced in the research. By manipulating these conditions, the results of this research will produce optimization of parameter k with a higher level of accuracy than previous research.

The data in the study amounted to 100,000 data and the data came from a website with the address www.kaggle.com which contains a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c levels, and blood glucose levels.

The following is information regarding the dataset used:

Dataset Name:
diabetes.

About Datasets:
The file contains the patient's medical and demographic data along with their diabetes status, both positive and negative.

Dataset Contents:
• gender: Gender refers to an individual's biological sex, which can impact their susceptibility to diabetes.
• age: Age is an important factor because diabetes is more often diagnosed in older adults.
• hypertension: Hypertension is a medical condition in which the blood pressure in the arteries is constantly elevated.
• heart_disease: Heart disease is another medical condition associated with an increased risk of diabetes.
• smoking_history: A history of smoking is also considered a risk factor for diabetes and may worsen related complications.
• bmi: BMI (Body Mass Index) is a measure of body fat based on weight and height. Higher BMI values are associated with higher risk.
• hbA1c_level: a measure of a person's average blood sugar levels over the last 2-3 months.
• blood_glucose_level : Blood glucose level refers to the amount of glucose in the bloodstream at a given time.
• diabetes: predicted target variable, with a value of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes.

The system proposed in this research is divided into two stages, namely the training and testing stages. In the training stage it is used to create the best classification model that can be used to optimize the k parameters in the k-NN algorithm, while in the testing stage it is used to classify whether you have diabetes or not.

In building a model, there are 3 main stages, namely data understanding, data preparation, modeling. Each stage can be seen in the following image 3.1:

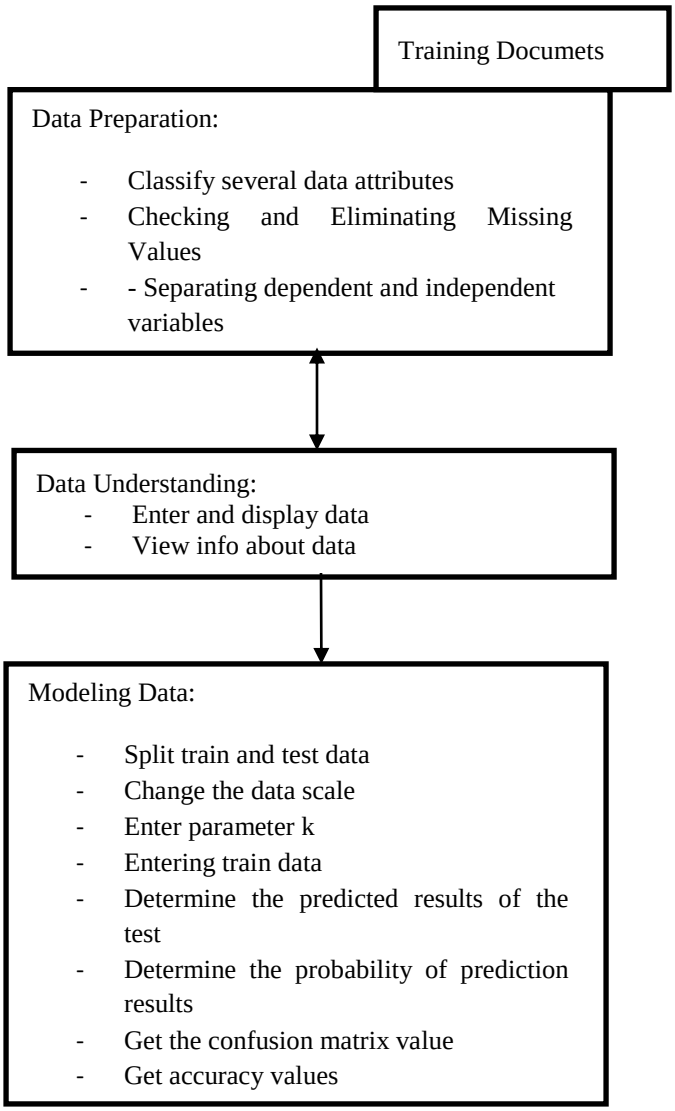

Figure 3.1. Development of research models

4. Results and Discussion

Test results for the parameter k with values 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41 , 43, 45, 47, and 49 get the confusion matrix value and presented in Table 4.1 as follows

Table 4.1. Confusion Matrix Value

Nilai Parameter K	Confusion Matrix			
	TP	FN	FP	TN
3	18040	211	626	1123
5	18105	146	678	1071
7	18170	109	681	1068
9	18162	89	696	1053
11	18181	70	696	1053
13	18192	59	711	1038
15	18192	59	721	1028
17	18191	60	730	1019
19	18201	50	741	1008
21	18204	47	747	1002
23	18211	40	746	1003
25	18216	35	749	1000
27	18218	33	748	1001
29	18219	32	751	998
31	18228	23	760	989
33	18227	24	762	987
35	18230	21	767	982
37	18229	22	771	978
39	18233	18	772	977
41	18232	19	773	976
43	18231	20	774	975
45	18233	18	779	970
47	18231	20	782	967
49	18232	19	784	965

Test results for the parameter k with values 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41 , 43, 45, 47, and 49 get the accuracy value and presented in Table 4.2 as follows:

Table 4.2. Accuracy Value

Parameter Value K	Accuracy
3	0.95815
5	0.9588
7	0.9605
9	0.96075
11	0.9617
13	0.9615
15	0.961
17	0.9605
19	0.96045
21	0.9603
23	0.9607
25	0.9608
27	0.96095
29	0.96085
31	0.96085
33	0.9607
35	0.9606
37	0.96035
39	0.9605
41	0.9604
43	0.9603
45	0.96015
47	0.9599
49	0.95985

From the table above, the research results show that the optimization parameters K=3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41 , 43, 45, 47, and 49 and getting the research results for optimizing the value of k = 11 having the highest accuracy of 0.9617.

5. Conclusions and recommendations

Conclusions for this research include:

The highest parameter optimization is produced by the highest k value from the values k=3,5,7,9,11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39 , 41, 43, 45, 47, and 49 which is 0.9617

Suggestions for this research include:

For now the research only focuses on optimizing the k parameters in the k-NN algorithm, then it is possible to optimize using a different algorithm.

Bibliography

[1] Indrayanti, et al. Optimasi Parameter K Pada Algoritma K-Nearest Neighbour Untuk Klasifikasi Penyakit Diabetes Mellitus. ISBN: 978-602-1180-50-1, Prosiding SNATIF Ke-4 Tahun 2017

[2] Nurkhozin, Agus, et al. Komparasi Hasil Penyakit Diabeter Mellitus Menggunakan Jaringan Syaraf Tiruan BackPropagation dan Learning Vector Quantization. Prosiding Seminar Nasional Penelitian, Pendidikan dan Penerapan MIPA, Fakultas MIPA, Universitas Negeri Yogyakarta, 14 Mei 2011.

[3] Pebrianti, Lidia, et al. Implementasi Metode Adaboost untuk Mengoptimasi Klasifikasi Penyakit Diabetes dengan Algoritma Naïve Bayes. E-ISSN: 25541-5735, P-ISSN: 2502-5724 Volume 7, No. 2, Agustus 2022

[4] Robbani, Alif Abqori, et al. Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma C4.5. Scientific Student Journal for Information, Technology and Science ISSN: 2715-2766 Vol. III No: 1, Januari 2022

[5] Infodatin 2020 Diabetes Melitus.pdf. diakses pada 15 Juli 2023 dari https://www.kemkes.go.id/downloads/resources/download/pusdatin/infodatin/

[6] HR, Siti Zulaikhah, et al. Optimasi Algoritma K-Nearest Neighbor (KNN) Dengan Normalisasi dan Seleksi Fitur Untuk Klasifikasi Penyakit Liver. JATI (Jurnal Mahasiswa Teknik Informatika) Vol. 6 No. 2, September 2022

[7] Hanafi, M. Habib, et al. Optimasi Algoritma K-Nearest Neighbor untuk Klasifikasi Tingkat Kematangan Buah Alpukat Berdasarkan Warna. IT Journal Research and Development (ITJRD) Vol.4, No.1, Agustus 2019, E-ISSN : 2528-4053 | P-ISSN : 2528-4061 DOI : 10.25299/itjrd.2019.vol4(1).2477

[8] Banjarsari, Mutiara Ayu, et al. Penerapan K-Optimal Pada Algoritma Knn untuk Prediksi Kelulusan Tepat Waktu Mahasiswa Program Studi Ilmu Komputer Fmipa Unlam Berdasarkan IP Sampai Dengan Semester 4. Kumpulan jurnaL Ilmu Komputer (KLIK) Volume 02, No.02 September 2015 ISSN: 2406-7857

[9] Prasetio, Rizki Tri. Seleksi Fitur dan Optimasi parameter k-NN berbasis algoritma genetika pada dataset medis. JURNAL RESPONSIF, Vol. 2 No.2 Agustus 2020, pp. 213~221 E-ISSN: 2685-6964

[10] Binabar, Sattriedi Wahyu dan Ivandari. Optimasi Parameter K pada Algoritma KNN untuk Deteksi Penyakit Kanker Payudara. IC-Tech Volume XII No. 2 Oktober 2017 http://jurnal.stmik-wp.ac.id

[11] Alkaromi, M. Adib. Optimasi Parameter K Pada Algoritma KNN untuk klasifikasi Heregistrssi Mahasiswa. Jurnal IC-Tech Vol.X No.1 April 2015 www.jurnal.stmik-wp.ac.id

[12] Ulya, Syaiful, et al. Optimasi Parameter K Pada Algoritma K-NN Untuk Klasifikasi Prioritas Bantuan Pembangunan Desa. Techno.COM, Vol. 20, No. 1, Februari 2021: 83-96

[13] Saputra, Arie Yandi dan Primadasa, Yogi. Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbour. Techno.COM, Vol. 17, No. 4, November 2018 : 395-403

[14] Nikmatun, Inna Alvi dan Waspada, Indra. Implementasi data mining untuk klasifikasi masa studi mahasiswa menggunakan algoritma k-nearest neighbor. Jurnal SIMETRIS, Vol. 10 No. 2 November 2019 P-ISSN: 2252-4983, E-ISSN: 2549-3108

[15] Rizal, Muftahul, et al. OPTIMASI ALGORITMA NAÏVE BAYES MENGGUNAKAN FORWARD SELECTION UNTUK KLASIFIKASI PENYAKIT GINJAL KRONIS. NARATIF : Jurnal Ilmiah Nasional Riset Aplikasi dan Teknik Informatika Vol. 05 No. 01 Juni 2023 P-ISSN: 2656-7377 || E-ISSN: 2714-8467

[16] Amalia, Hilda. OPTIMASI NEURAL NETWORK MENGGUNAKAN GENETIC ALGORITHM UNTUK PREDIKSI PENYAKIT DIABETES. PARADIGMA Vol. XVII. No.2 September 2015

[17] Azzahrah, Diah Siti Fatimah dan Alamsyah. Klasifikasi Penyakit Diabetes Menggunakan Algoritma K-Nearest Neighbor Optimasi K-Fold Cross Validation. Seminar Nasional Ilmu Komputer (SNIK 2022) - Semarang, 19 Oktober 2022 ISSN: 2614-1205

[18] Ariani, Ardina dan Samsuryadi. Klasifikasi Penyakit Ginjal Kronis menggunakan K-Nearest Neighbor. Prosiding Annual Research Seminar 2019 Computer Science and ICT ISBN : 978-979-587-846-9 Vol.5 No.1

[19] Handayanna, Frisma, et al. PREDIKSI PENYAKIT DIABETES MENGGUNAKAN NAIVE BAYES DENGAN OPTIMASI PARAMETER MENGGUNAKAN ALGORITMA GENETIKA. Konferensi Nasional Ilmu Sosial & Teknologi (KNiST) Maret 2017, pp. 71~76

[20] Maskuri, Muhammad Naja, et al. Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Memprediksi Penyakit Stroke. Jurnal Ilmiah Intech : Information Technology Journal of

UMUS Vol.4, No.1, Mei 2022 pp. 130~140
[21] Al Karomi, M. Adib. OPTIMASI PARAMETER K PADA ALGORITMA KNN UNTUK KLASIFIKASI HEREGISTRASI MAHASISWA. Jurnal IC-Tech Vol.X No.1 April 2015, www.jurnal.stmik-wp.ac.id
[22] Pangestu, Raka Aji, et al. Optimasi Nilai k Pada Algoritma K-Nearest Neighbor Untuk Klasifikasi Pasien Covid-19 Yang Membutuhkan Ruangan ICU. JURNAL INOVTEK POLBENG - SERI INFORMATIKA, VOL. 7, NO. 1, 2022
[23] Christobel, Y. Angeline and Subramanian, Suresh. Predicting Diseases Using Optimal 'K' Value in KNearest Neighbor for Human Safe. Journal of Green Engineering (JGE) Volume-10, Issue-5, May 2020
[24] Wijanarto dan Puspitasari, Rhatna. Optimasi Algoritma Klasifikasi Biner dengan Tuning Parameter pada Penyakit Diabetes Mellitus. JURNAL EKSPLORA INFORMATIKA
[25] Silalahi, Arina Prima, et al. SUPERVISED LEARNING METODE K-NEAREST NEIGHBOR UNTUK PREDIKSI DIABETES PADA WANITA. METHOMIKA: Jurnal Manajemen Informatika & Komputerisasi Akuntansi Vol. 7 No. 1 (April 2023), ISSN: 2598-8565 (media cetak), ISSN: 2620-4339 (media online)
[26] Hana, Fida Maisa dan Wahyudin, Widya Cholid. OPTIMASI PARAMETER A LGORITMA DECISION T REE C4.5 PADA KLASIFIKASI B LOGGER PROFESSIONAL. urnal Ilmu Komputer dan Matemtika Vol. 4 No. 2 (2023) 77-83 | 77