DIABETES PREDICTION AS FOOD RECOMMENDATION SYSTEM USING CONTENT-BASED FILTERING BASED ON ANDROID

Oleh : Sufyan Hanif Ariyana[1], Parsumo Rahardjo[2], Sukamto[3]
Jurusan Teknik Elektro[1,2,3], Politeknik Negeri Semarang[1,2,3]
E-mail : sufyan.43321025@mhs.polines.ac.id[1] , parsumo@polines.ac.id[2],
sukamto@polines.ac.id[3]

Abstrak

Diabetes tipe 2 merupakan kondisi kronis dengan prevalensi global yang terus meningkat, yang dipengaruhi oleh Body Mass Index(BMI) yang tidak normal dan kebiasaan makan yang buruk. Penelitian ini bertujuan untuk mengembangkan sistem berbasis machine learning untuk memprediksi risiko diabetes dan memberikan rekomendasi diet yang dipersonalisasi berdasarkan BMI dan hasil prediksi. Metodologi penelitian ini dirancang sendiri dan diilustrasikan melalui diagram alir yang terdiri dari enam tahap: pengumpulan data, persiapan data, pemodelan, evaluasi, berbasis aturan, dan penerapan. Data terkait diabetes dikumpulkan dari Rumah Sakit Roemani (2020-2024), sementara data makanan dikumpulkan melalui web scraping dari situs web FatSecret. Model prediksi menggunakan algoritma Support Vector Classifier (SVC) dan mencapai akurasi 97,77%. Metode pemfilteran berbasis konten digunakan untuk rekomendasi makanan, menghasilkan Mean Absolute Error (MAE) sebesar 0,9362. Sistem ini digunakan sebagai aplikasi Android, menawarkan saran makanan yang dipersonalisasi untuk membantu pengguna mengontrol kebiasaan makan dan menurunkan risiko diabetes tipe 2.

Kata kunci: *Diabetes tipe 2, Machine Learning, SVC, Content-Based Filtering, Aplikasi Android*

Abstract

Type 2 diabetes is a chronic condition with rising global prevalence, influenced by abnormal Body Mass Index (BMI) and poor dietary habits. This study aims to develop a machine learning-based system for predicting diabetes risk and providing personalized dietary recommendations based on BMI and prediction results. The methodology is self-designed and illustrated through a flowchart consisting of six stages: data collection, data preparation, modeling, evaluation, rule-based, and deployment. Diabetes-related data was collected from Roemani Hospital (2020–2024), while food data was gathered through web scraping from the FatSecret website. The prediction model uses the Support Vector Classifier (SVC) algorithm and achieves an accuracy of 97.77%. A content-based filtering method is used for food recommendation, producing a Mean Absolute Error (MAE) of 0.9362. The system is deployed as an Android application, offering personalized food suggestions to help users control dietary habits and lower their risk of type 2 diabetes.

Keywords: *Type 2 Diabetes, Machine Learning, SVC, Content-Based Filtering, Android Application*

1. Introduction

Type 2 diabetes is a medical condition characterized by chronically high blood sugar levels. Based on data from the International Diabetes Federation (IDF), it is estimated that 463 million people worldwide - or about 9.3% of the global population aged 20 to 79 - will have diabetes in 2019. This number is expected to continue to rise, reaching 700 million by 2045 and 578 million by 2030 [1]. This disease can occur due to various factors, one of which is poor eating habits. Poor eating habits, such as consuming high-calorie and low-nutrient foods, can disrupt the body's ability to effectively regulate blood sugar levels [2].

A poor diet begins with the selection of unhealthy foods, such as the consumption of high-calorie, low-nutrient foods and processed foods rich in sugar. Research by Ullin Saranianingsi analyzes the impact of poor and irregular eating patterns on blood sugar levels in diabetes patients. The results show a significant relationship between poor eating habits and increased blood sugar levels, which could potentially worsen the patient's condition [3].

To address the issue, in this era of rapid technology, we can use Machine Learning (ML) [4] to assist in selecting the appropriate food recommendations according to the user's condition. In the research by Luqyana in 2024, ML was implemented with the KNN model in a food recommendation system based on nutritional needs using Content-Based Filtering, resulting in an RMSE of 18.75, which means the food recommendation model is in accordance with the input profile [5]. It is intended that the population will obtain dietary recommendations that are in line with nutritional requirements and reduce the risk of developing more severe diabetes by employing the Content-Based Filtering methodology.

2. Method

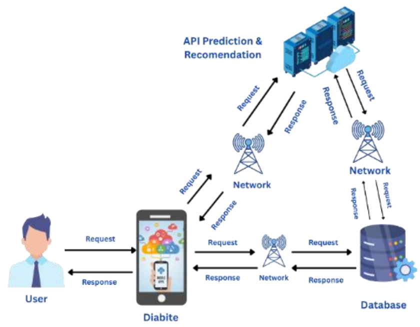

Figure 1 Architecture diagram of the application

In Figure 1 is the process of using the diabite application by users. Users will input their profile into the application and send it. The API model will run and send the user's condition results and food recommendations to the user's application. The user's condition results and food recommendations will automatically enter the database and become the user's history in the application. All these processes can be carried out if connected to the internet.

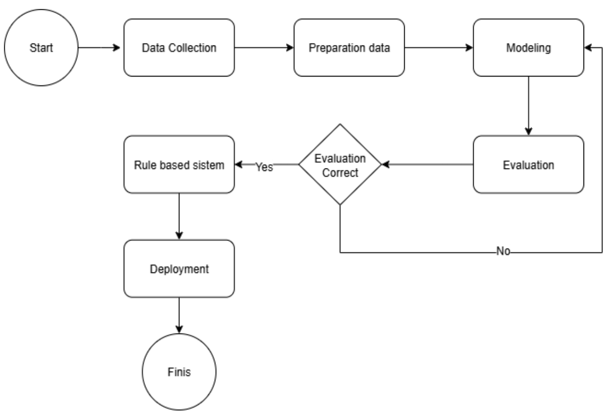

Figure 2 Flowchart method for system development

The methodology employed in this study to develop a machine learning model and meal recommendations consists of six steps, as shown in Figure 2. This is the method's explanation:

2.1 Data Collection

Data collection is the process of gathering relevant data and exploring it to obtain an overview and assess the quality of the data. In this research, there are 2 types of data needed, namely diabetes patient data and food nutritional content data.

Diabetes prediction data used for diabetes prediction includes information on blood sugar levels, gender, age, height, weight, and previous medical history such as diabetes and hypertension. This data was collected from the hospital Roemani system database during the period 2020-2024, totaling 554. Here is an explanation of the attributes present in the dataset. The features used in the predictions in this study can be seen in Table 1.

Table 1 Dataset Diabetes

Field	Explain
Blood Sugar/gds	Blood sugar taken from a gluco meter that can be tested at any time.
Age	Age of patient
Height	Patient's height in cm
Weight	Patient's weight in kg
Gender	Patient’s gender
History of hypertension	Patient's medical history if 1 has experienced/has blood pressure over 140 mmHg and a headache [6] while 0 never experienced
History of diabetes	Previous history of the patient's disease if 1 has experienced, 0 if not experienced
BMI	Body Mass Index of the patient's height and weight.
Diabetes	The result of predicting diabetes is 1 if diabetic and 0 if not diabetic.

The food data was obtained from web scraping using the Chrome extension on the

FatSecret website totaling 121 foods with a serving size of 100 grams, and of course, only healthy foods were selected. Here is an explanation of the attributes present in the dataset as shown in Table 2.

Table 2 Dataset Food

Field	Explain
Name	Name Food
Types	Type of food (vegetables, fruit, meat fish, bread, soup, beans)
Calories	Calorie content of food in kcal
Carbohydrates	Carbohydrate content of food in grams
Fat	Fat content of food in grams
Protein	Protein content of food in grams

2.2 Preparation Data

The obtained data will go through several important stages, starting from cleaning (removing incomplete, duplicate, or incorrect data), transformation (adjusting the format and structure of the data), normalization (standardizing the scale of values), to feature selection (choosing the most relevant variables) to ensure the quality and readiness of the data before analysis or modeling.

2.3 Modeling

Diabetes prediction for diabetes prediction modeling uses one of the best models in terms of accuracy and not overfitting. The models used to conduct this research are supervised learning algorithms, namely decision tree, logistic regression, random forest, support vector machine, and K-nearest neighbors. There are 2 scenarios for making diabetes prediction models seen in Table 3.

Table 3 Training and Testing Data Scenario

Skenario	Data Size	
	Traing	Testing
1	70%	30%
2	80%	20%

The food recommendation system is designed to model a food recommendation system using a content-based filtering method that applies Euclidean distance for distance measurement. This approach was selected because, in the context of food nutritional content, it is important to understand the quantitative differences between various foods.

2.4 Evaluation Model

The model used must perform model performance testing. In prediction, there are several performance testing metrics, such as precision, recall, F1 score, and learning curve. While the food recommendation system uses Mean Average Precision (MAP).

Precision measures the ability of the model to correctly identify an object as a true positive among all declared positive results (both true and false). Formula (1) is used to calculate precision[7].

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

Recall is the capacity of the model to identify every positive instance that actually exists in the data. Recall thus measures the number of positive cases (false negatives) that the model fails to detect. The following recall formula is addressed in formula (2).

$$Recall = \frac{TP}{TP+FP} \quad (2)$$

Accuracy is the ratio of the total number of correct predictions and the total number of predictions. Accuracy is calculated by formula (3).

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (3)$$

A learning curve is a visual tool used to illustrate how model performance changes as the amount of training data increases. By plotting the model score (such as accuracy, precision, or recall) against the size of the training data, the learning curve helps to identify whether the model is underfitting or overfitting.

A confusion matrix is a table that summarizes the performance of the classification model by comparing the predicted label with the actual label. The confusion matrix will display the amount of data classified in the correct (true positive and true negative) and incorrect (false positive and false negative) classes[8].

MAP is a standard for evaluating model performance and provides an overall picture of the recommender model's performance by averaging the Average Precision (AP) values at each recommendation ranking level addressed in formula (3). MAP values that are close to 1 indicate that the recommendation is good [9]. Formula (4) is used to calculate MAP.

$$MAP = \frac{1}{m} \sum_{i=1}^{m} APi \qquad (4)$$

2.5 Rule Based System

Rule-Based System (RBS) is an artificial intelligence system that relies on a set of explicit rules to make decisions or provide recommendations. These rules are usually in the form of IF-THEN, where certain conditions (if met) trigger specific actions or outputs [10]. In this research, rule-based methods are used as a connector between two models to obtain food preferences, as shown in Table 4.

Table 4 Rule Based System

Patient Condition	**Parameters**	**BMI Criteria**	**Recommended Food Types**	**Food Selection Criteria**
Possible Diabetes	Carbohydrate < 45 grams/serving (130 grams/day divided by 3) [11]	< 18,5	Vegetables, Fruit, Nuts, Fish & Seafood, Eggs [12]	Highest calories on the food list [13]
		18,5–24,9		Calories close to TDEE [14]
		> 24,9		Lowest calories and lowest carbs on the food list [13]
Negative Diabetes	-	< 18,5	Vegetables, Fruit, Nuts, Fish & Seafood, Eggs, Soups, Meat	Highest calories on the food list

Patient Condition	Parameters	BMI Criteria	Recommended Food Types	Food Selection Criteria
		18,5–24,9		Calories close to TDEE [14]
		> 24,9		Lowest calories and lowest carbs on the food list [13]

2.6 Deployment

The model that has been made will be implemented in the form of an application using Flutter as a programming framework [15] the application is called Diabite. By combining the 2 best models of diabetes prediction models with food recommendation models. The following is a flowchart on the application later seen in Figure 3.

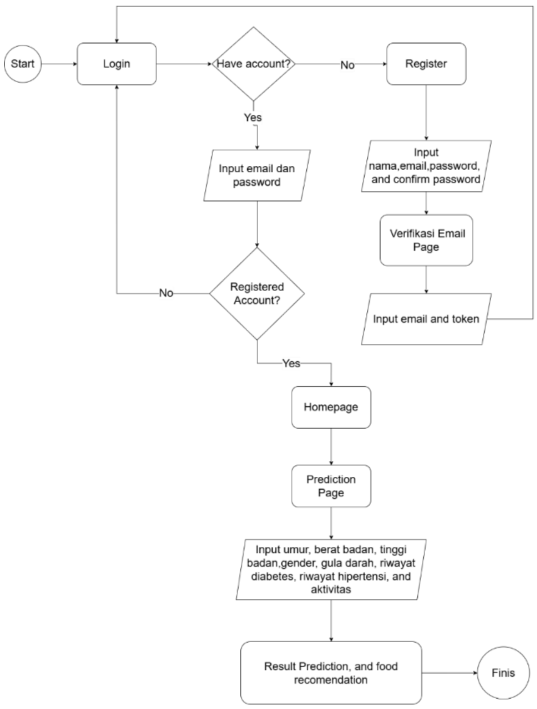

Figure 3 Flowchart aplication

The program flow is addressed in Figure 3. The program begins with login if you have an account and registration if you want to register. If you choose login, of course there is an account validation process; if it is appropriate to eat, you will proceed to the home page. The next page is the home page. On the home page there is an add button that will direct to the diabetes prediction page. On the diabetes prediction page, users can enter age, gender, weight, height, blood sugar, and activity; after that, select the save button, and it will direct them to the prediction results page and food recommendation results.

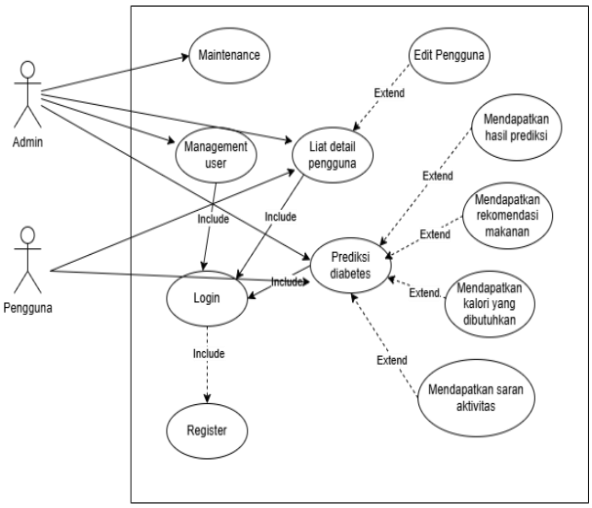

Figure 4 Usecase diagram

3. Result And Discussion

The following results and discussion of models that have been made for diabetes prediction and food recommendations are explained as follows:

3.1 Diabetes Prediction Model

In training, the data model is divided into two scenarios: the first scenario is 70% data and 30% testing, and the second scenario uses 80% training data and 20% testing data. Training data aims to build and optimize the model, while testing data aims to evaluate the performance of the model that has been made. This division aims to allow the model to learn from most data while being tested on data that has never been seen before, so that the evaluation results reflect the model's ability to generalize to new data accurately and validly. In this thesis, the algorithm model is compared with the logistic regression, decision tree, KNN, random forest, and support vector machine models. The comparison results can be seen in Table 5 for 70% training and 30% testing and table 6 for 80% training and 20% testing.

Table 5 Algorithm Comparison Results 70% Training 30% Testing

Model Name	Precision	Recall	F1-Score	Testing Accuracy	Training Accuracy
Logistic Regression	96%	87%	91%	90,13%	92,63%
Decision Tree	100%	99%	99%	99,34%	100%
Knn	94%	93%	93%	92,76%	96,03%
Random Forest	99%	100%	99%	99,34%	100%
Support Vector Machine	97%	97%	97%	96,05%	97.45%

Table 6 Algorithm Comparison Results 80% Training 20% Testing

Model Name	Precision	Recall	F1-Score	Testing Accuracy	Training Accuracy
Logistic Regression	96%	92%	94%	93.07%	94,01%
Decision Tree	100%	100%	100%	100%	100%
Knn	93%	93%	93%	92,08%	96,29%
Random Forest	98%	100%	99%	99,01%	100%
Support Vector Machine	98%	97%	97%	97,03%	97,77%

After testing several models that can be seen in table 5 and table 6, the support vector machine model with 80% training data and 20% testing data has good results, not overfitting and good accuracy, namely precision 97%, recall 97%, f1-score 97%, testing accuracy 97.03% and training accuracy 97.77%. The next process to validate model performance is using learning curves. The following results of the SVC learning curves model performance are shown in Figure 5:

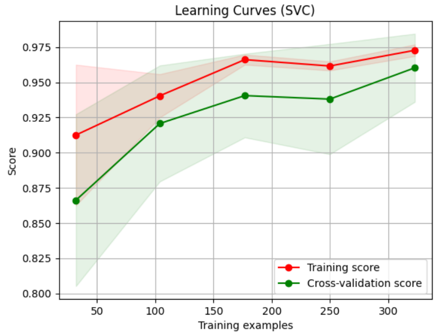

Figure 5 Learning curvers SVC

Based on Figure 5, the SVC learning curves of 80% training data and 20% testing data show the training score with a red line, and the green line shows the cross-validation score. It can be seen that the training score and cross-validation score both increase as the amount of training data increases. The training score continues to rise and reaches an almost maximum value, while the cross-validation score also shows a steady increase, although it does not rise as much as the training score. The difference between these two scores remains relatively small, indicating that even though the model is highly trained on the training data, it can still generalize well to the testing data or data that has not been seen before, without showing any indication of overfitting. This indicates that the model can learn well from the training data and retain the ability to perform effectively on unfamiliar data. The final evaluation of the confusion matrix for the diabetes prediction model is shown in table 7.

Table 7 Confusion Matrix

	Prediction Class 0	Prediction Class 1
Class 0	True Negative (TN) = 40	False Positive (FP) = 1
Class 1	False Negative (FN) = 2	True Positive (TP) = 58

With these results, the model shows SVM performance is good and can be implemented in applications as a diabetes prediction model. The last process is testing the SVM model with the following hospital data, the results can be seen in Table 8.

Table 8 Diabetes Prediction Model Testing Results

Gender	Age	History of hypertension	Height	Weight	Blood sugar	History of Diabetes	Diabetes	Result
Male	71	Yes	150	86	294	No	Positive	Correct
Female	60	Yes	155	50	221	No	Positive	Correct
Female	37	No	151	51	110	No	Negative	Correct
Male	42	Yes	152	98	360	No	Positive	Correct
Female	30	No	152	53	162	No	Negative	Correct

3.2 Prediction Model of Food Recommendation System

The test chosen is Euclidean distance because in the context of the nutritional content of food, it is needed to know how much quantitative difference there is between foods with mAP (mean average precision) = 0.9362. The following are the results of the food recommendation system that has similarities with lettuce shown in Table 9.

Table 9 Lettuce Recommendation Results from 4 Food Types

Name	Types	Calories	Fat	Carbohydrates	Protein	*Distance Skor*
Seledri	Vegetables	14	0,17	2,97	0,69	0,026839
Selada daun hijau	Vegetables	15	0,15	2,79	1,36	0,059107
Stroberi	Fruits	32	0,30	7,68	0,67	0,324229
Jeruk bali	Fruits	32	0,10	8,08	0,63	0,340304

Buncis	Peanuts	31	0,12	7,13	1,82	0,313064
Kancang panjang hijau	Peanuts	31	0,12	7,13	1,82	0,313064
Tiram	Fish & *Seafood*	68	2,46	3,91	7,05	1,079015
Cumi-cumi	Fish & *Seafood*	91	1,37	3,05	15,45	2,013325

3.3 Deployment

Validation system applications using a combination of 2 SVM models and content-based filtering with rule-based as a connector for the 2 models. Types of food recommendations for diabetics. The following test results of the overall food recommendation system that has been deployed in the application can be seen in Figure 6.

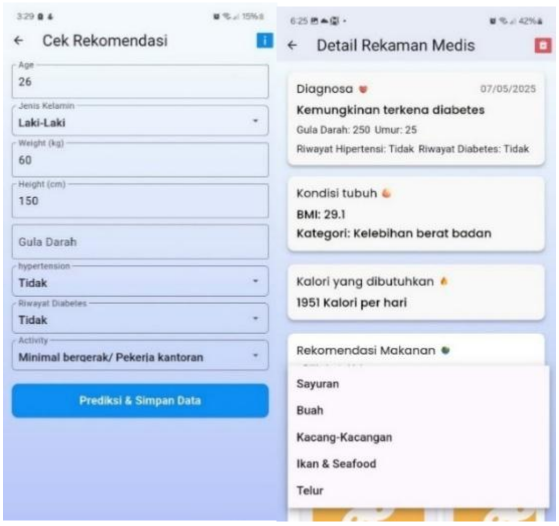

Figure 6 Implementation In The Application (a) and (b)

Figure 6 shows the test results of the food recommendation system obtained from diabetes prediction through the rule base process well as the final recommendation using the content-based filtering method. Figure (a) is the process of inputting the user's condition and Figure (b) is the result of the test. For users who are potentially diabetic, the system only provides four types of food recommendations in accordance with the rules set in the rule base.

Table 10 Diabetes Prediction Results and Rule-Based Food Preferences

Gender	**Age**	**History of hypertension**	**Height**	**Weight**	**Gula darah**	**History of diabetes**	**Diabetes**	**Food Preferences**
Male	25	No	170	60	150	0	No	Corn flakes

Female	35	Yes	170	65	210	1	Yes	Kuning telur matang
Female	37	No	151	51	110	0	No	Selada
Male	42	Yes	152	98	360	1	Yes	Selada

Table 10 is the result of diabetes prediction and the goal is to get food preferences as food recommendations using content-based filtering. The following content-based filtering results can be seen in table 11.

Table 11 Food Recommendation System Results

Diabetes Prediction	Food preferences	Recommendation result
No	Corn flakes	Ubi Rambat, wortel, prune kering, jeruk, cumi, tiram, kacang merah tanpa garam, kacang merah dengan garam, telur, telur rebus, daging sapi rebus dengan kentang dan sayuran, daging sapi rebus dengan kentang dan wortel, hati bebek, daging bebek, roti panggang pita, roti pita
Yes	Kuning telur matang	Zaitun hijau, ubi rambat, alpukat, rasberi, ikan sarden, ikan salmon, kacang panggang, kacang merah, telur bebek, dan telur rebus
No	Selada	Seledri, sekada daun hijau, stroberi, jeruk bali, tiram, cumi,buncis, kacang Panjang hijau, putih telur, putih telur matang, sup ayam, sup tomat, paha ayam, daging paha ayam, oatmeal, dan oatmeal instan
Yes	Selada	Seledri, sekada daun hijau, stroberi, jeruk

		bali, tiram, cumi,buncis, kacang Panjang hijau, putih telur, dan putih telur matang.

Table 11 is the result of the food recommendation system taken from 1 food preference and diabetes prediction, which determines the type of food suggested. If the patient's condition is diabetic, then 5 types of food will be recommended, namely vegetables, fish and seafood, eggs, fruit, and nuts, while if not diabetic.

4. Conclusion

A food recommendation system using the content-based filtering model eudilance distance algorithm with MAE 0.9362 and a diabetes prediction system using the SVC algorithm with an accuracy of 97.77% in comparison to the decision tree algorithm model, logistic regression, random forest, and k-nearest neighbors can be inferred from this study. The food recommendation system does not use cosine similarity because it requires the amount of quantitative difference between foods, not only proportional similarity, when discussing food nutritional content. The primary function of this program is to offer meal suggestions based on the user's health profile and condition forecast, along with diabetes-related predictions.

Reference

[1] A. Azriful, Y. Adnan, E. Bujawati, S. Alam, and N. Nildawati, “Mengungkap Fakta Faktor Risiko Diabetes Melitus Di Indonesia,” *Media Penelitian dan Pengembangan Kesehatan*, vol. 34, no. 4, pp. 814–823, Nov. 2024, doi: 10.34011/jmp2k.v34i4.1988.

[2] D. Jauhanita, A. Sriatmi, and M. I. Kartasurya, “Manajemen Diabetes Melitus pada Remaja: Evaluasi Terhadap Pendekatan Nutrisi dan Intervensi Gizi dalam Tinjauan Sistematis,” *MAHESA : Malahayati Health Student Journal*, vol. 4, no. 5, pp. 1946–1964, May 2024, doi: 10.33024/mahesa.v4i5.14460.

[3] U. Saranianingsi, S. Rammang, E. Hidayat, I. Keperawatan, and U. Widya Nusantara, “Hubungan Pola Makan dan Aktivitas Fisik terhadap Peningkatan Kadar Gula Darah Pasien Diabetes Melitus Tipe 2 di Puskesmas Biromaru,” *https://jptam.org*, vol. 9, pp. 2894–2899, 2025.

[4] Dakhaz Mustafa Abdullah and Adnan Mohsin Abdulazeez, “Machine Learning Applications based on SVM Classification: A Review,” *Qubahan Academic Journal*, vol. 3, no. 4, pp. 206–218, Nov. 2021, doi: 10.48161/Issn.2709-8206.

[5] Luqyana Zakiya Almas, Yuliana Susanti, and Sri Sulistijowati Handajani, “Penerapan Algoritma K-Nearest Neighbors dalam Sistem Rekomendasi Makanan Berdasarkan Kebutuhan Nutrisi dengan Content-Based Filtering,” *Statistika*, vol. 24, no. 1, pp. 115–122, May 2024, doi: 10.29313/statistika.v24i1.3558.

[6] R. Zainuddin and P. Labdullah, “Efektivitas Isometric Handgrip Exercise dalam Menurunkan Tekanan Darah pada Pasien Hipertensi Effectiveness Of Isometric Handgrip Exercise In Reducing Blood Pressure In Hypertension Patients,” *Jurnal Ilmiah Kesehatan Sandi Husada*, 2020, doi: 10.35816/jiskh.v10i2.364.

[7] Hovi Sohibul Wafa, Asep Id Hadiana, and Fajri Rakhmat Umbara, “Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM) INFORMASI ARTIKEL ABSTRAK,” *INFORMATICS AND DIGITAL EXPERT (INDEX)*, vol. 4, no. 1, pp. 40–45, 2022, [Online]. Available: https://e-journal.unper.ac.id/index.php/informatics

[8] N. Maulidah, R. Supriyadi, D. Y. Utami, F. N. Hasan, A. Fauzi, and A. Christian, “Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes,” *Indonesian Journal on Software Engineering (IJSE)*, vol. 7, no. 1, pp. 63–68, 2021, [Online]. Available: http://ejournal.bsi.ac.id/ejurnal/index.php/ijse63

[9] Evidently AI Team, “Mean Average Precision (MAP) in ranking and recommendations,” Evidently AI.

Accessed: Jul. 01, 2025. [Online]. Available: https://www.evidentlyai.com/ranking-metrics/mean-average-precision-map#how-to-calculate-map

[10] Ardhi Supratman, Indarmawan Nugroho, and Rifki Dwi Kurniawan, “Penerapan Metode Rule Based System Untuk Menentukan Jenis Tanaman,” *INNOVATIVE: Journal Of Social Science Research*, vol. 4, no. 2, pp. 7879–7890, 2024.

[11] F. F. Zakiyah, V. Indrawati, S. Sulandjari, and S. A. Pratama, “Asupan karbohidrat, serat, dan vitamin D dengan kadar glukosa darah pada pasien rawat inap diabetes mellitus,” *Jurnal Gizi Klinik Indonesia*, vol. 20, no. 1, p. 21, Jul. 2023, doi: 10.22146/ijcn.83275.

[12] A. Susilowati, B. Rachmat, R. Ayu Larasati, P. Penelitian dan Pengembangan Upaya Kesehatan Masyarakat, B. Penelitian dan Pengembangan Kesehatan, and K. R. Kesehatan Jl Percetakan, “Hubungan Pola Konsumsi Serat Dengan Kontrol Glikemik Pada Diabetes Tipe 2 [T2d] Di Kecamatan Bogor Tengah (Relationship Of Fiber Consumption Patterns To Glycemic Control In Type 2 Diabetes [T2d] In Central Bogor Sub-District),” *Penel Gizi Makan*, no. 1, 2020.

[13] I. B. Nafisah, “Analisis Hubungan Antara Indeks Massa Tubuh (Bmi), Ketidakpuasan Tubuh Dan Lingkungan Sosial Terhadap Perilaku Diet Tidak Sehat Pada Gadis Remaja,” *JURNAL KESEHATAN TAMBUSAI*, vol. 5, no. 2, 2024.

[14] V. Fresinsya and A. Qoiriah, “Meal-Planning Berbasis Status Gizi dengan Metode Klasifikasi K-nearest neighbors (KNN) untuk Pasien Obesitas,” *Journal of Informatics and Computer Science*, vol. 06, 2024.

[15] N. Sofi and R. Dharmawan, “Perancangan Aplikasi Bengkel Csm Berbasis Android Menggunakan Framework Flutter (Bahasa Dart),” *JTS*, vol. 1, no. 2, 2022.