Jurnal Media Teknik Elektro dan Komputer
Vol. XX, No.XX (January - June, 2024)
E-ISSN: XXXX-XXXX

Expert System for Detecting Potential Obesity in the Environment of the State Islamic University of North Sumatra Using Naïve Bayes Classifier

Hardiansyah Putra[1], Inun Suryani Harahap[2]

[1] STMIK Royal Kisaran; hardiansyahputrahsb@gmail.com
[2] Universitas Islam Negeri Sumatera Utara; inunsuryaniharahap@gmail.com

ABSTRACT

This study aims to classify the potential for obesity among students in the environment of UIN SU Medan using the Naïve Bayes Classification method. The prevalence of obesity has become a serious problem in various countries, with significant contributions from critical periods such as prenatal, infancy, adiposity rebound, and adolescence. The use of technology has become widespread in various layers of society, making knowledge and preventive efforts crucial in addressing obesity. Lifestyle changes due to increased time-consuming activities can affect the health of the academic community. Therefore, the application of technology is essential in addressing obesity from adolescence to adulthood. The Naïve Bayes Classification method has been found to be suitable for classifying the likelihood of obesity among the academic community in the UIN SU Medan environment. However, it is important to optimize the training and testing classification process by storing the results in memory (database) to enhance the classification process.

Keywords: Expert Systems, Classification, Obesity, Naïve Bayes Classifier

Corresponding Author:
Hardiansyah Putra
STMIK Royal Kisaran; hardiansyahputrahsb@gmail.com

1. INTRODUCTION

The introduction should briefly explain the research in a broad context and show why it is important. This section should explain the study's purpose and its significance. Cite key publications and review the current state of research. If necessary, include hypotheses that are crucial, important, or even points of controversy, along with the pros and cons of different opinions and perspectives. Finally, state the purpose of the study and its main findings. Try to make the introduction understandable to readers outside the research field.

The use of technology has spread to various levels of society, where all activities now rely on it as a support, from light to heavy tasks. One example is the application of technology in the health sector (Maryani & Irmayansyah, 2023). A balanced proportion of body weight and height is considered good for a person's health. According to Brunner Simanjuntak, D. (2019), obesity has become a serious problem in various countries in the world, with a prevalence rate of 30% in the United States and 22%

in the UK in 2001 (Sufi et al., 2023). There are four critical periods that contribute to obesity, namely the prenatal period, infancy, adiposity rebound period, and adolescence. As many as 30% of obesity cases in adolescence tend to continue into adulthood as persistent obesity (Simanjuntak & Sindar, 2019). Being overweight, even just a few kilograms, can pose significant health risks. Therefore, knowledge and prevention efforts are crucial in managing obesity from adolescence to adulthood (Delima & Prasetio, 2021).

Determination of obesity levels can be done by measuring waist and hip circumference, with the ratio of waist and hip circumference being a very significant factor related to the risk of coronary heart disease (CHD) (Erwis et al., 2022). We used data mining techniques in the analysis of this research data. Data mining is a technique in data processing that identifies relationships in data that are unknown to users and presents them in an easy-to-understand format (Harliana & Anggraini, 2023).

Decision-making can use the results of this relationship identification as a basis. Specific tasks such as description, estimation, prediction, classification, clustering, and association can group data mining. The method of classification groups new data into predetermined categories. We use the classification algorithm to categorize the level of obesity in the UINSU Medan environment. We carry out identification based on weight and height data, as well as unhealthy lifestThis study utilizes the questionnaire results as research material, classifying the determined data from 300 samples in the UIN SU Medan Environment (Aldisa et al., 2022).

Research criteria are necessary for evaluating a person's health. Health in this context includes the lifestyle of activities carried out by members of the academic community in the UIN SU Medan environment. There are changes in the lifestyle of students when they are involved in activities that require more time, and this can affect the health of the entire academic community in the UIN SU Medan environment. Based on the results of the questionnaire that has been distributed to the academic community in the UIN SU Medan environment, a classification will be carried out on the level of possibility of obesity in information systems students caused by an unhealthy lifestyle. According to Akbar's research (2021), the factors causing obesity in adolescents are complex, including lack of sleep, spending time on static activities, and unhealthy eating patterns (Akbar, 2021). The study's results revealed that the average obesity rate among adolescents was 10.5% among the students under investigation. Based on previous research using the Na've Bayes Classification algorithm method, the results indicate that this method is highly suitable for use in this particular case. Therefore, in classifying the possibility of obesity in academics in the UIN SU Medan Environment, the Naïve Bayes Classification method can be applied (Asmarani et al., 2022).

(Asmarani et al., 2022) research on the classification of nutritional status yielded results of 93.2% using the Naive Bayes Classification method, which is widely regarded as having good performance. However, (Majid et al., 2020) point out that the training and testing process of this method, if not stored in memory (database), is not optimal. Therefore, to optimize the classification process, we recommend storing the classification results in memory (a database).

2. LITERATURE REVIEW

2.1. Data Mining

(Majid et al., 2020) refer to data mining as knowledge discovery in databases (KDD), an activity that involves collecting and utilizing historical data to identify patterns or relationships in large datasets. Future decision-making can benefit from the results of this data mining process. Therefore, data mining rarely uses the term pattern recognition separately, as it forms an integral part of the concept (Majid et al., 2020). In general, Enisa and Mulyani (2021) divide training methods in data mining techniques into two main approaches (Genisa & Mulyana, 2021).

1. Unsupervised Learning
   a. We apply this method without any training or teacher guidance.
   b. The label of the data, called the teacher, is not needed in the application of this method.
2. Supervised Learning
   a. This method involves training and guidance from the trainer.
   b. To find the decision function, separation function, or regression function, several examples of data with output or labels are used during the training process.

Some consider data mining as a synonym for knowledge discovery from data or the commonly used KDD, while others see data mining as an important step in the knowledge discovery process (Alpiansah & Ramdhani, 2023). The knowledge discovery process includes several stages, namely:

1. Data cleaning aims to remove noise and inconsistent data.
2. Data integration involves combining multiple data sources.
3. Data Selection: the process of selecting relevant data for the analysis task from the database.
4. Data Transformation: Perform summary aggregation operations to transform and consolidate data into a form suitable for mining.
5. Data mining: an essential process that applies intelligent methods to extract data patterns
6. Pattern Evaluation: Identifying truly interesting patterns, representing knowledge based on established interest criteria.
7. Knowledge Presentation: This involves the use of visualization and presentation techniques to present mining results to users (Rahmah et al., 2021).

Figure 1 illustrates the various stages and mechanisms that make up the data mining process.

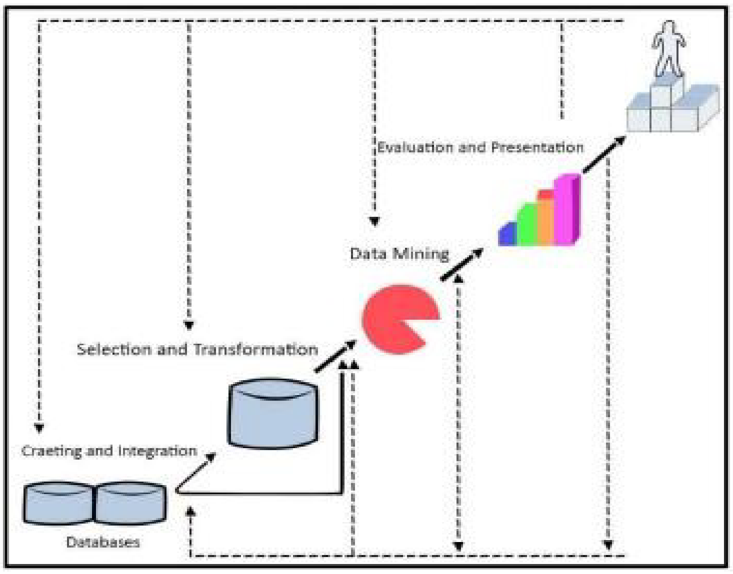


Figure 1. Data Mining Process

2.2. *The Naïve Bayes classification algorithm*

Bayes decision theory is one of the most important statistical approaches in pattern recognition. It measures the costs and benefits of different classification choices by looking at the chances and costs of each one (Arieska, 2024) Additionally, Bayesian classification is capable of predicting the probability of class membership, as demonstrated by Bayes' theorem, which shares classification capabilities with decision trees and neural networks. (Nugroho & Prambodo, 2022) have demonstrated the high accuracy and speed of Bayesian classification in databases with large data. Additionally, a variety of supporting devices, including RapidMiner, can facilitate the data mining process. Various techniques in machine learning, data mining, text mining, and predictive analytics environments use RapidMiner as a tool (Bani & Asruddin, 2022).

3. METHODS

Determining the appropriate method to apply is the first step. Next, we conduct a literature study on the proposed research topic. The next step entails gathering the required data, which we do by conducting a survey with a questionnaire and selecting a random sample. We carry out data selection after data collection to determine which data is suitable for use and which is not. Moreover, RapidMiner, a tool, undergoes a transformation process on the data. The next step is to apply the Naïve Bayes Classification method to perform calculations. We use the same tool for testing and evaluation after the calculations are complete to ensure the accuracy and validity of the results. The last step in this research methodology is to display the final results in graphical form, providing a clear visual depiction of the findings and conclusions of the study. Figure 2 illustrates the research methodology in the following diagram.

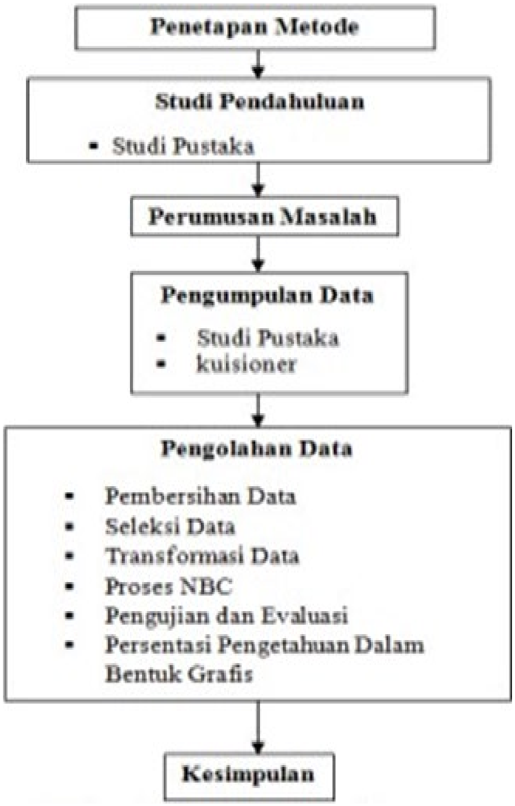


Figure 2. Research Stages

4. RESULTS AND DISCUSSION

In its development, the application of an expert system for detecting diabetes in the environment of the State Islamic University of North Sumatra requires data that will be used in its classification. The dataset in question is obtained by distributing questionnaires in the form of several questions that refer to the parameters of diabetes or its symptoms. The questionnaire that has been made by the researcher is then given to several respondents who are willing to fill it out. The total dataset obtained in this study is 299 rows. Previously, the dataset had gone through the data preprocessing stage, in the form of cleaning data from missing data (missing value), vague or fuzzy data (noise), and information failure. Figure 3 displays the dataset under consideration.

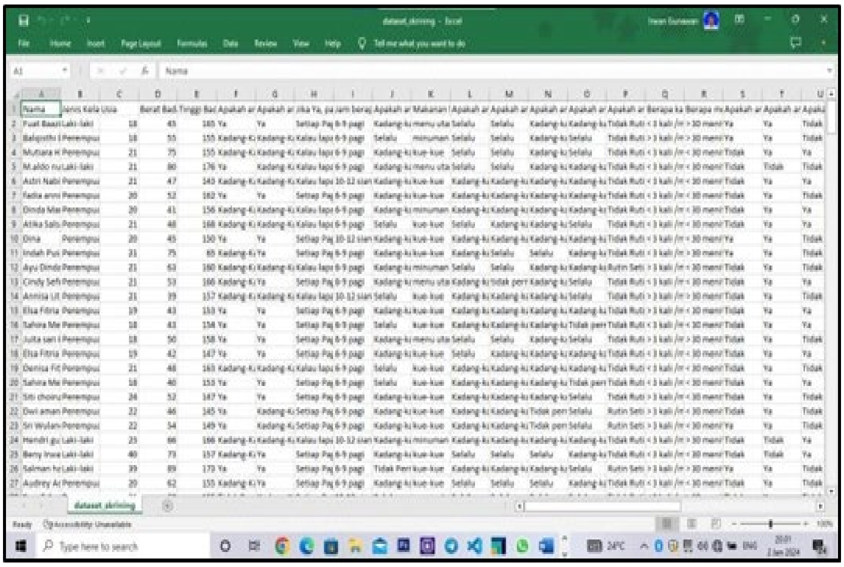


Figure 3. Description of Research Dataset

Next, we will enter the cleaned dataset into the RapidMiner application, assigning a rule to one column that serves as a label or target, specifically the diabetes and non-diabetes classes. Figure 4 illustrates this process, and Figure 5 shows that there were previously 299 datasets entered into Rapidminer.

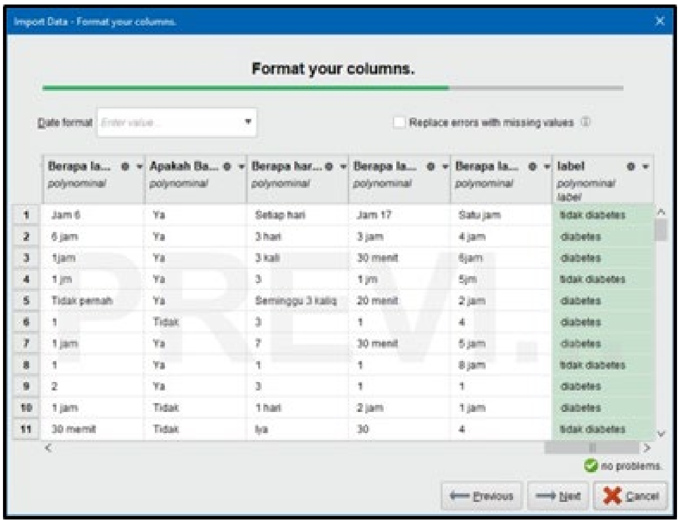
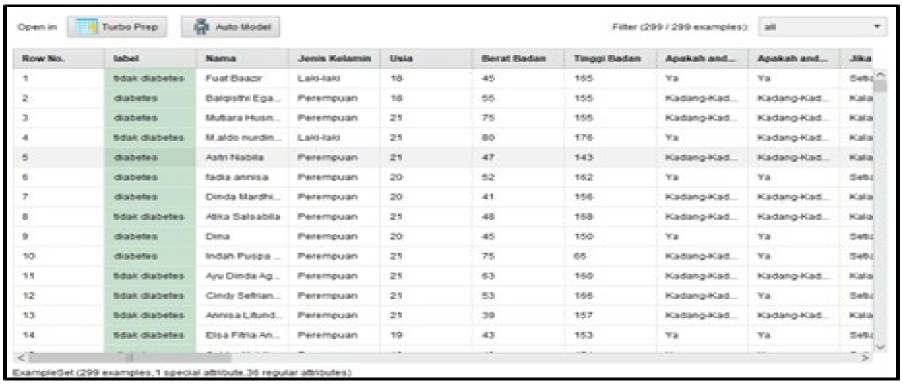


Figure 4. Labeling the Dataset



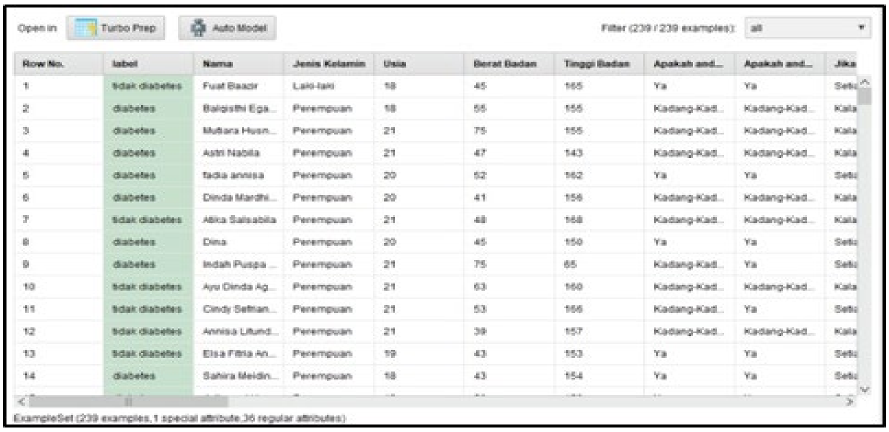
Open in | Turbo Prep | Auto Model | Filter (299 / 299 examples): all

Row No.	label	Nama	Jenis Kelamin	Usia	Berat Badan	Tinggi Badan	Apakah and...	Apakah and...	Jika
1	tidak diabetes	Fuat Baazir	Laki-laki	18	45	165	Ya	Ya	Seti
2	diabetes	Balqisthi Ega...	Perempuan	18	55	155	Kadang-Kad...	Kadang-Kad...	Kala
3	diabetes	Mutiara Husn...	Perempuan	21	75	155	Kadang-Kad...	Kadang-Kad...	Kala
4	tidak diabetes	M.aldo nurdin...	Laki-laki	21	80	176	Ya	Kadang-Kad...	Kala
5	diabetes	Astri Nabila	Perempuan	21	47	143	Kadang-Kad...	Kadang-Kad...	Kala
6	diabetes	fadia annisa	Perempuan	20	52	162	Ya	Ya	Seti
7	diabetes	Dinda Mardhi...	Perempuan	20	41	156	Kadang-Kad...	Kadang-Kad...	Kala
8	tidak diabetes	Atika Salsabila	Perempuan	21	48	168	Kadang-Kad...	Kadang-Kad...	Kala
9	diabetes	Dina	Perempuan	20	45	150	Ya	Ya	Seti
10	diabetes	Indah Puspa ...	Perempuan	21	75	65	Kadang-Kad...	Ya	Seti
11	tidak diabetes	Ayu Dinda Ag...	Perempuan	21	63	160	Kadang-Kad...	Kadang-Kad...	Kala
12	tidak diabetes	Cindy Seftian...	Perempuan	21	53	166	Kadang-Kad...	Ya	Seti
13	tidak diabetes	Annisa Litund...	Perempuan	21	39	157	Kadang-Kad...	Kadang-Kad...	Kala
14	tidak diabetes	Elsa Fitria An...	Perempuan	19	43	153	Ya	Ya	Seti

ExampleSet (299 examples, 1 special attribute, 36 regular attributes)

Figure 5. Number of Diabetes Screening Datasets

In classification, the terms "training data" and "test data" refer to the data used to train a classification model, and "test data" refers to the data used to test the model. This study utilizes diabetes screening datasets. We will separate the training data by 80%, or 239 rows, and the test data by 20%, or 60 rows, in the classification standard. The machine learning process necessitates a large amount of training data for the computer learning process. Figure 5 displays the training data in question, while Figure 6 displays the test data.

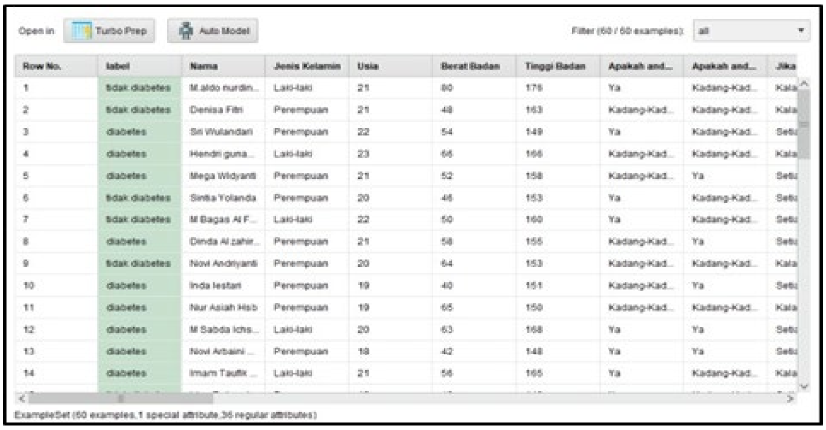


Open in | Turbo Prep | Auto Model | Filter (239 / 239 examples): all

Row No.	label	Nama	Jenis Kelamin	Usia	Berat Badan	Tinggi Badan	Apakah and...	Apakah and...	Jika
1	tidak diabetes	Fuat Baazir	Laki-laki	18	45	165	Ya	Ya	Seti
2	diabetes	Balqisthi Ega...	Perempuan	18	55	155	Kadang-Kad...	Kadang-Kad...	Kala
3	diabetes	Mutiara Husn...	Perempuan	21	75	155	Kadang-Kad...	Kadang-Kad...	Kala
4	diabetes	Astri Nabila	Perempuan	21	47	143	Kadang-Kad...	Kadang-Kad...	Kala
5	diabetes	fadia annisa	Perempuan	20	52	162	Ya	Ya	Seti
6	diabetes	Dinda Mardhi...	Perempuan	20	41	156	Kadang-Kad...	Kadang-Kad...	Kala
7	tidak diabetes	Atika Salsabila	Perempuan	21	48	168	Kadang-Kad...	Kadang-Kad...	Kala
8	diabetes	Dina	Perempuan	20	45	150	Ya	Ya	Seti
9	diabetes	Indah Puspa ...	Perempuan	21	75	65	Kadang-Kad...	Ya	Seti
10	tidak diabetes	Ayu Dinda Ag...	Perempuan	21	63	160	Kadang-Kad...	Kadang-Kad...	Kala
11	tidak diabetes	Cindy Seftian...	Perempuan	21	53	166	Kadang-Kad...	Ya	Seti
12	tidak diabetes	Annisa Litund...	Perempuan	21	39	157	Kadang-Kad...	Kadang-Kad...	Kala
13	tidak diabetes	Elsa Fitria An...	Perempuan	19	43	153	Ya	Ya	Seti
14	diabetes	Sahira Meidin...	Perempuan	18	43	154	Ya	Ya	Seti

ExampleSet (239 examples, 1 special attribute, 36 regular attributes)

Figure 6. Diabetes Screening Training Datasets



Open in | Turbo Prep | Auto Model | Filter (60 / 60 examples): all

Row No.	label	Nama	Jenis Kelamin	Usia	Berat Badan	Tinggi Badan	Apakah and...	Apakah and...	Jika
1	tidak diabetes	M.aldo nurdin...	Laki-laki	21	80	176	Ya	Kadang-Kad...	Kala
2	tidak diabetes	Denisa Fitri	Perempuan	21	48	163	Kadang-Kad...	Kadang-Kad...	Kala
3	diabetes	Sri Wulandari	Perempuan	22	54	149	Ya	Kadang-Kad...	Seti
4	diabetes	Hendri guna...	Laki-laki	23	66	166	Kadang-Kad...	Kadang-Kad...	Kala
5	diabetes	Mega Widyanti	Perempuan	21	52	158	Kadang-Kad...	Ya	Seti
6	tidak diabetes	Sintia Yolanda	Perempuan	20	46	153	Ya	Kadang-Kad...	Seti
7	tidak diabetes	M Bagas Al F...	Laki-laki	22	50	160	Ya	Kadang-Kad...	Seti
8	diabetes	Dinda Al zahir...	Perempuan	21	58	155	Kadang-Kad...	Ya	Seti
9	tidak diabetes	Novi Andriyanti	Perempuan	20	64	153	Kadang-Kad...	Kadang-Kad...	Kala
10	diabetes	Inda lestari	Perempuan	19	40	151	Kadang-Kad...	Ya	Seti
11	diabetes	Nur Asiah Hsb	Perempuan	19	65	150	Kadang-Kad...	Kadang-Kad...	Kala
12	diabetes	M Sabda Ichs...	Laki-laki	20	63	168	Ya	Ya	Seti
13	diabetes	Novi Arbaini ...	Perempuan	18	42	148	Ya	Ya	Seti
14	diabetes	Imam Taufik ...	Laki-laki	21	56	165	Ya	Kadang-Kad...	Kala

ExampleSet (60 examples, 1 special attribute, 36 regular attributes)

Figure 7. Diabetes Screening Testing Datasets

After the datasets go through the data cleaning stage (preprocessing) and the stage of separating training data and test data (split data), the next step is to design a classification model to detect diabetes using the RapidMiner supporting application. An operator called naïve bayes is used to make a classification model. An operator called apply model is used to test the model made from training data with test data (testing). Finally, Figure 8 shows the design of the model that was meant to be used to see how well the classification model made using the performance operator works.

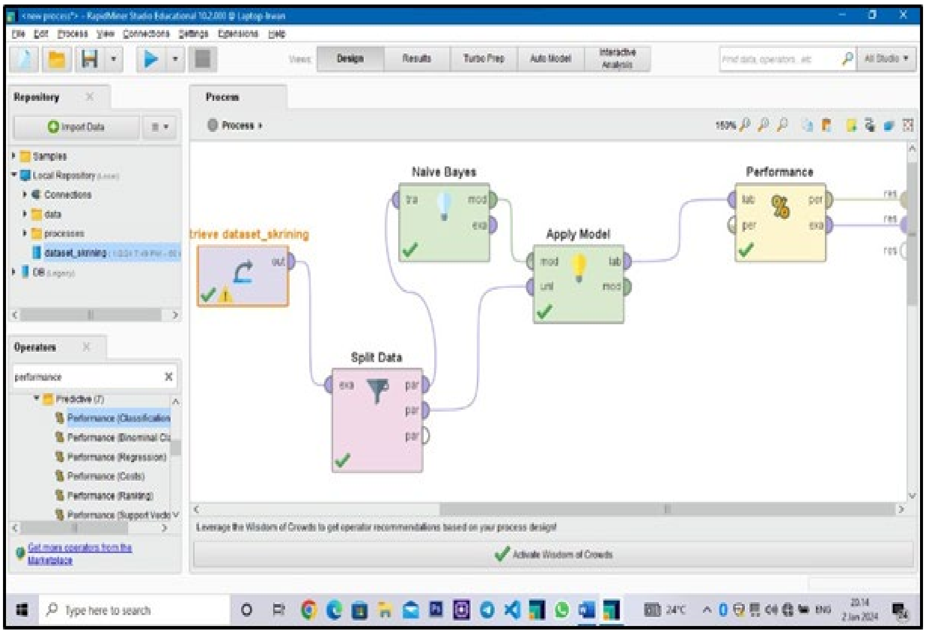


Figure 8. Creating a Naïve Bayes Classifier Model

Figure 7 depicts the creation and testing of the classification model. The results of the prediction accuracy with the target or label are obtained by comparing the training confidence and testing confidence values based on 60 rows of test data. The intended prediction results can be seen in Figure 9.

Row No.	label	prediction(la...	confidence(t...	confidence(...	Jenis Kelamin	Usia	Berat Badan	Tinggi Badan	Apal
1	tidak diabetes	diabetes	0.027	0.973	Laki-laki	21	80	176	Ya
2	tidak diabetes	diabetes	0.001	0.999	Perempuan	21	48	163	Kad
3	diabetes	diabetes	0.053	0.947	Perempuan	22	54	149	Ya
4	diabetes	tidak diabetes	0.609	0.391	Laki-laki	23	66	166	Kad
5	diabetes	diabetes	0.050	0.950	Perempuan	21	52	158	Kad
6	tidak diabetes	diabetes	0.020	0.980	Perempuan	20	46	153	Ya
7	tidak diabetes	tidak diabetes	0.999	0.001	Laki-laki	22	50	160	Ya
8	diabetes	diabetes	0.135	0.865	Perempuan	21	58	155	Kad
9	tidak diabetes	tidak diabetes	0.862	0.138	Perempuan	20	64	153	Kad
10	diabetes	diabetes	0.000	1.000	Perempuan	19	40	151	Kad
11	diabetes	tidak diabetes	0.999	0.001	Perempuan	19	65	150	Kad
12	diabetes	diabetes	0.000	1.000	Laki-laki	20	63	168	Ya
13	diabetes	diabetes	0.022	0.978	Perempuan	18	42	148	Ya
14	diabetes	diabetes	0.000	1.000	Laki-laki	21	56	165	Ya

ExampleSet (60 examples,4 special attributes,35 regular attributes)

Figure 9. Results of Diabetes Disease Classification on Datasets

Next, to find out the level of accuracy based on the prediction results in Figure 8, cross validation will be carried out, this stage will calculate the number of predictions that match the failed predictions, the cross validation in question can be seen in Figure 10.

accuracy: 63.33%

	true tidak diabetes	true diabetes	class precision
pred. tidak diabetes	4	11	26.67%
pred. diabetes	11	34	75.56%
class recall	26.67%	75.56%	

Figure 10. Cross Validation Performance Results from Datasets

Based on the classification results to detect diabetes in the problems mentioned earlier using diabetes screening datasets in the State Islamic University of North Sumatra Medan with a total of 299 rows of data, the accuracy results were 63,33%; this is quite good in detecting diabetes in the web system that will be developed later.

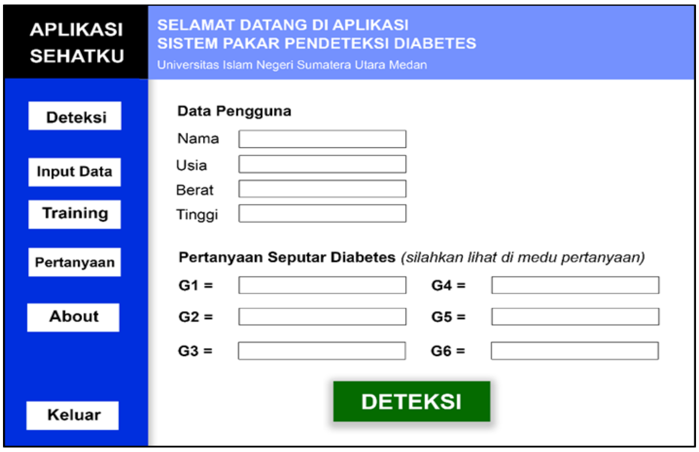


Figure 11. View of the Diabetes Disease Detection Expert System Application

By implementing a previously created classification model using the PHP programming language and a MySQL database, we create an expert system application.

5. CONCLUSION

The research, which utilized the diabetes screening dataset to predict obesity, yielded a classification model capable of detecting obesity with an accuracy rate of 63.33% when tested on 60 data samples. In addition, the application created using the PHP programming language and the MySQL database can be applied to build and design the intended expert system. The hope is that the results of this study can help the community learn how to independently detect diabetes. The researcher provides several suggestions for researchers and further research so that they can update the datasets used to have good data patterns by paying attention to the preprocessing stages and parameters used. In addition, a way is needed to increase the level of accuracy, such as changing the questionnaire question parameters to determine diabetes to be more relevant and accurate in detecting.

REFERENCES

Akbar, R. (2021). PERANCANGAN SISTEM PAKAR DIAGNOSA POTENSI SERANGAN STROKE BERBASIS JAVA. *Jurnal Perencanaan, Sains, Teknologi Dan Komputer*, *4*(1), 213–231.

Aldisa, R. T., Alfarisi, S., & Abdullah, M. A. (2022). Penerapan Metode Naïve Bayes Dalam Mendiagnosa Penyakit Leptospirosis. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 521–526. https://doi.org/10.47065/josyc.v3i4.2205

Alpiansah, A. B., & Ramdhani, Y. (2023). Optimasi Fitur dengan Forward Selection pada Estimasi Tingkat Obesitas menggunakan Random Forest Feature Optimization with Forward Selection on Obesity Rate Estimation using Random Forest. *SISTEMASI: Jurnal Sistem Informasi*, *12*(September), 860–873.

Arieska, R. (2024). Edukasi Gizi Seimbang Pada Remaja Di Wilayah BTN Pagesangan Pepabri Melalui Penggunaan Aplikasi Deteksi Dini Dan Edukasi Potensi Obesitas. *ALKHIDMAH: Jurnal Pengabdian Dan Kemitraan Masyarakat*, *2*(1), 171–176.

Asmarani, A., Permana, I., Putri, A., Wijaya, M. R., Rasywir, E., Meisak, D., & Pratama, Y. (2022). Implementasi Algoritma K-Nearst Neighbor Untuk Memprediksi Penyakit Diabetes. *Jurnal Informatika Dan Rekayasa Komputer(JAKAKOM)*, *2*(2), 231–239. https://doi.org/10.33998/jakakom.2022.2.2.110

Bani, A. U., & Asruddin, A. (2022). Pendeteksian Penyakit Mulut dan Kuku Pada Sapi dengan Menerapkan Metode Naïve Bayes. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 264–268. https://doi.org/10.47065/josyc.v3i4.1934

Delima, D. P., & Prasetio, R. T. (2021). Sistem Pakar Diagnosa Komplikasi Obesitas Pada Remaja Menggunakan Metode Certainty Factor. *EProsiding Sistem Informasi (POTENSI)*, *2*(1), 51–60.

Erwis, F., Suherdi, D., Pranata, A., & Nasyuha, A. H. (2022). Penerapan Metode Hybrid Case Base Pada Sistem Pakar Diagnosa Penyakit Obesitas. *Jurnal Media Informatika Budidarma*, *6*(1), 378. https://doi.org/10.30865/mib.v6i1.3491

Genisa, L., & Mulyana, D. I. (2021). Implementasi Penerapan Metode C4.5 dan Naïve Bayes Dalam Tingkat Kelulusan Akreditasi Lembaga PAUD Pada Badan Akreditasi Nasional. *Jurnal Media Informatika Budidarma*, *5*(4), 1595. https://doi.org/10.30865/mib.v5i4.3267

Harliana, & Anggraini, D. (2023). Penerapan Algoritma Naïve Bayes Pada Klasifikasi Status Gizi Balita di Posyandu Desa Kalitengah. *Jurnal Informatika Komputer, Bisnis Dan Manajemen*, *21*(2), 38–45. https://doi.org/10.61805/fahma.v21i2.16

Majid, A. M., Nuraeni, R., Anshor, A. H., Informatika, T., & Bangsa, U. P. (2020). *Jurnal Teknologi Pelita Bangsa Penerapan Data Mining Untuk Penentuan Status Gizi Balita Menggunakan Algoritma Naive Bayes (Studi Kasus Posyandu Seruni XII Kelurahan Karangraharja)*. *12*(1), 1–7.

Maryani, I., & Irmayansyah, I. (2023). Penerapan Algoritma Naïve Bayes Untuk Penentuan Diagnosa Obesitas Pada Peserta Sosialisasi Deteksi Dini Penyakit Tidak Menular (PTM). *TeknoIS : Jurnal Ilmiah Teknologi Informasi Dan Sains*, *13*(2), 234–248. https://doi.org/10.36350/jbs.v13i2.200

Nugroho, F., & Prambodo, Y. L. (2022). Pendeteksian Penyakit Limfadenopati dengan Menerapkan Metode Naive Bayes. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 199–204. https://doi.org/10.47065/josyc.v3i4.1997

Rahmah, S. A., Voutama, A., & Sobur, S. (2021). Sistem Pakar Diagnosis Obesitas Pada Orang Dewasa Menggunakan Metode Backward Chaining. *INTECOMS: Journal of Information Technology and Computer Science*, *4*(2), 169–177. https://doi.org/10.31539/intecoms.v4i2.2538

Simanjuntak, D., & Sindar, A. (2019). Sistem Pakar Deteksi Gizi Buruk Balita dengan Metode Naïve Bayes Classifier. *Jurnal Inkofar **, *1*(2), 2581–2920.

Sufi, H., Utomo, D. W., & Darmawati, G. (2023). Sistem Pakar Rekomendasi Menu Makanan Sehat Penderita Penyakit dengan Metode Forward Chaining. *Jurnal KomtekInfo*, *10*, 8–14. https://doi.org/10.35134/komtekinfo.v10i1.320