Semesta Teknika
Vol. 27, No. 1, 28-36, Mei 2024
DOI: https://doi.org/10.18196/st.v27i1.21008

Naive Bayes for Diabetes Prediction: Developing a Classification Model for Risk Identification in Specific Populations

Ahmad Zaki Arrayyan[1*] , Hendra Setiawan[1], Karisma Trinanda Putra[2]
[1]Departement of Electrical Engineering, Faculty of Industrial Technology, Universitas Islam Indonesia; [2]Departement of Electrical Engineering, Faculty of Engineering, Universitas Muhammadiyah Yogyakarta

[1]Jl. Kaliurang, Ngemplak, Sleman, Yogyakarta; [2]Jl. Brawijaya, Tamantirto, Kasihan, Bantul, Yogyakarta

[*] Corresponding author email: 22925007@students.uii.ac.id

Keywords:

Diabetes; naïve bayes; performance matrix

Abstract

Depending on persuasive statistics, the increasing prevalence of diabetes worldwide is a huge challenge for individuals, families, and nations. According to International Diabetes Federation (IDF) projections, the number of adults with diabetes is expected to rise by an astounding 46% by 2045, to reach 783 million, or one in eight. In response to this growing concern, this research explores the implementation of the Naive Bayes algorithm for predicting diabetes, employing comprehensive data cleansing and randomization techniques. A systematic evaluation of the model's performance is conducted using several training and testing split ratios (65:35, 75:25, 85:15). The outcome showed that the model performed best at the 65:35 split ratio, with accuracy reaching its maximum of 88.16%, precision 0.883, recall 0.881, and f1-score 0.882.

INTRODUCTION

The escalating global impact of diabetes poses a significant challenge for individuals, families, and nations, as evidenced by compelling statistics. IDF Diabetes Atlas (2021) reports that 10.5% of adults between the ages of 20 and 79 have diabetes, almost half of them being ignorant of their illness (Al-Mohaithef et al., 2022). Projections from the International Diabetes Federation (IDF) suggest a staggering 46% increase by 2045, resulting in 783 million adults, or 1 in 8, grappling with diabetes (Arokiasamy et al., 2021a). Type 2 diabetes accounts for over 90% of cases, influenced by a complex interplay of socio-economic, demographic, environmental, and genetic factors (Arokiasamy et al., 2021b). Factors such as urbanization, an aging population, declining physical activity, and rising overweight and obesity rates contribute significantly to the rise in type 2 diabetes(Patil & Gothankar, 2019).

Nevertheless, proactive measures can mitigate the impact of diabetes, emphasizing preventive actions for type 2 diabetes and early diagnosis and care for all forms of the condition (Association, 2019). These interventions are crucial to helping individuals manage the condition, potentially avoiding or delaying complications. While diabetes cannot be completely cured, it can be effectively managed (Khursheed et al., 2019). Insulin is used to treat type 1 diabetes. It is available in different forms, including rapid-acting, short-acting, intermediate-acting, and long-acting insulin (Janež et al., 2020). The differences in the types of insulin are in how quickly they take effect and how long they last. A balanced diet, oral medicines, and regular exercise can all help control type 2 diabetes (Magkos et al., 2020). "Insulin-Dependent Diabetes Mellitus" (IDDM), another name for Type 1 Diabetes Mellitus, is characterised by insufficient insulin synthesis by the pancreas (Esposito et al., 2019). Because their pancreas produces less insulin than normal, people with Type 1 diabetes need to take exogenous insulin injections to make up for this (Charley et al., 2023) . The hallmark of type 2 diabetes mellitus is the body's resistance to insulin due to abnormal responses of body cells to the hormone. The body may eventually run out of insulin as a result of this resistance. According to Sousa et al. (2021) it is also known as "Non-

Insulin-Dependent Diabetes Mellitus" (NIDDM) (Sousa et al., 2021). Sedentary lifestyles are a common cause of this kind of diabetes.

However, current methods for detecting diabetes involve the use of laboratory tests such as blood glucose and oral glucose tolerance (Kuo et al., 2021), the process of which is quite time-consuming. The urgent nature of this surge calls for innovative approaches, especially in early detection and prevention strategies. Identifying high-risk individuals has become crucial to curbing the diabetes epidemic.

Therefore, machine learning is considered an important feature of artificial intelligence that supports the development of computer systems with the ability to acquire knowledge from past experience quickly (Raschka et al., 2020). This study utilizes machine learning techniques, focusing on Naive Bayes classification to predict diabetes susceptibility in random individuals, with the aim of improving early detection and intervention.

In practical terms, Naive Bayes works by evaluating the probability of a particular instance to fall into a particular class based on the values of its features (Maswadi et al., 2021). The algorithm is "naive" in that it simplifies the calculations by assuming independence among the features (Chen et al., 2020), although this assumption does not always hold in real-world scenarios (Wickramasinghe & Kalutarage, 2021). Despite its simplicity and the assumptions made, Naive Bayes often performs well in various classification tasks, especially when dealing with large datasets (Abbas et al., 2019; Blanquero et al., 2021).

The essence of Naive Bayes lies in its ability to make predictions based on probability distributions, making it a powerful tool in the field of machine learning and classification (Hassan et al., 2023). Its simplicity, efficiency, and effectiveness make it a popular choice for various applications, in this case specifically medical diagnosis.

In the context of this study, it's important to note that the research leverages secondary data directly collected from patients at Diabetes Sylhet Hospital in Sylhet, Bangladesh. This data collection involved the use of questionnaires that were ethically approved by doctors, ensuring the ethical and methodological integrity of the research process. This utilization of real-world patient data adds a practical dimension to the application of Naive Bayes in medical diagnosis, offering insights into its performance in a specific healthcare context. The incorporation of such data enhances the relevance and applicability of the findings to real-world scenarios, contributing to the robustness of the study's conclusions.

RESEARCH METHOD

This study used secondary data collected using questionnaires directly from patients of Diabetes Sylhet Hospital in Sylhet, Bangladesh and approved by doctors. This study has the following methods:

Data Cleansing

Data cleaning is an important step in data preparation that involves examining the data set to identify and address missing values or inconsistent data. As explained from the UCI dataset, out of 520 data consisting of 17 attributes and 1 class, there were missing values. So this important process ensures the reliability and accuracy of the data by correcting discrepancies and improving its overall quality. It involves thorough verification to detect anomalies or irregularities that could jeopardize the integrity of the dataset.

Data Randomization

Data randomization is a crucial process involving the randomization of data to obtain a representative sample that accurately reflects the population. This method ensures that the selected sample is not biased and provides a fair and unbiased representation of the entire dataset. The goal is to eliminate any potential patterns or order in the data, promoting objectivity in subsequent analyses. The process of data randomization involves shuffling or rearranging the data points in a way that each observation has an equal chance of being selected. By doing so, the resulting sample is more likely to capture the diverse characteristics present in the overall population.

Naive Bayes Classification

The Naive Bayes algorithm is a classification technique that operates on the principles of probabilistic classification. Naive Bayes is a straightforward probabilistic classification method that calculates a set of probabilities by summing frequencies and combinations derived from a given data set. The theorem predicts future probabilities based on past experience and combined with the Naive assumption, which assumes independence between attributes. It is assumed in Naive Bayes classification that a feature's existence or absence in a class has no bearing on the presence of other features in that class. In the exploration of the dataset, different proportions for training and testing datasets namely 65:35, 75:25, and 85:15. These ratios represent the division of the dataset into training and testing subsets, with the first number indicating the percentage allocated for training and the second for testing. The rationale behind experimenting with different splits lies in finding an optimal balance between providing the algorithm with sufficient data to learn patterns and ensuring an adequate evaluation on unseen data. Previous works have often employed varying proportions to strike this balance, and comparing the results across different splits allows for a nuanced understanding of the algorithm's behavior under diverse conditions. It enables researchers to evaluate the model's ability to generalize beyond the training set and provides insights into potential overfitting or underfitting issues. Additionally, by scrutinizing the impact of these proportions on the Naive Bayes algorithm's performance, researchers can contribute to the broader discourse on best practices in dataset partitioning for classification tasks. The study further extends its investigation by calculating prior values from the training data and determining probability values for each variable, emphasizing a comprehensive evaluation of the algorithm's predictive capabilities under different training-test scenarios. Following this, the study delves into calculating the prior values of the training data and determining the probability values for each variable concerning its respective class through the application of the conditional probability formula.

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \quad (1)$$

$P(A|B)$ = is the probability of class A given the features B, $P(B|A)$ = is the probability of observing the features B given class A, $P(A)$ = is the prior probability of class A, and $P(B)$ = is the probability of observing the features B.

Classification Model Evaluation

Classification model evaluation is the process of assessing the performance and accuracy of a developed classification model. The primary goal is to understand how well the model classifies instances into different categories or classes. The evaluation is typically quantified using various metrics, and one common metric is accuracy, precision, recall, and f1-score. Accuracy is a measure that signifies the overall correctness of the model by considering both true positives and true negatives in relation to all instances. Precision is a metric focused on the accuracy of positive predictions, emphasizing the ratio of true positives to the total predicted positives. The ratio of true positives to all actual positives is called recall, or sensitivity, and it measures how well the model can detect positive occurrences. The f1-score is a balanced metric that takes into account both false positives and false negatives. It is calculated as the harmonic mean of precision and recall. Conversely, specificity focuses on how well the model detects negative examples; this is measured by the proportion of real negatives to all actual negatives.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F1 - score = \frac{2.Precision.Recall}{Precision+Recall} \quad (5)$$

In these formulas: *TP* = True Positives (correctly predicted positive instances), *TN* = True Negatives (correctly predicted negative instances), *FP* = False Positives (incorrectly predicted as positive instances), *FN* = False Negatives (incorrectly predicted as negative instances).

RESULTS AND DISCUSSION

This test aims to find the accuracy value and which performance is better between the division of testing data and training data using Naive Bayes in classifying data into predetermined classes. In this study, the data used amounted to 520 data, with 16 attributes namely age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, blurring of vision, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, obesity, and 1 class.

Table 1. Number of data training and data testing

No	Split of data training and testing	Data training	Data testing	Total data
1.	65:35	338	182	520
2.	75:25	390	130	520
3.	85:15	442	78	520

Referring to table 1 which is a representation of the amount of training and testing data division that will be carried out in this study.

Performance Matrix

Confusion matrix

Confusion matrix belongs to the category of performance matrix in the context of machine learning model evaluation. Confusion matrix provides a detailed description of the model's performance by comparing the model's prediction results with the actual value of the test data. the following are the results of the confusion matrix in this study from data training and testing 65:35, 75:25, 85:15

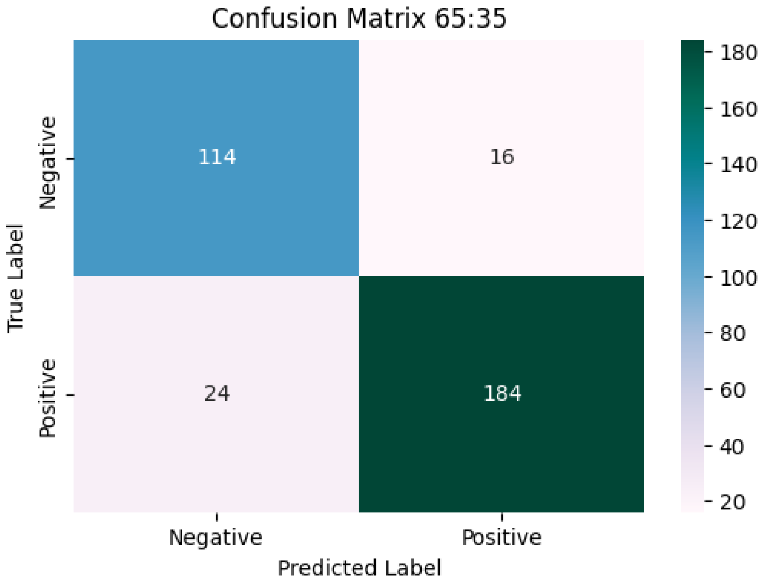

Figure 1. Confusion matrix data training and data testing 65:35

Based on the confusion matrix plot in figure 1, outcomes derived from the confusion matrix analysis reveal a detailed breakdown of the model's performance, showcasing 184 instances correctly identified as True Positives (TP), demonstrating the model's proficiency in accurately predicting positive cases. Simultaneously, the model achieved 114 instances of True Negatives (TN), correctly identifying negative cases. However, it also incurred 16 False Positives (FP), instances where the model inaccurately classified negative cases as positive, signifying the occurrence of Type I errors. Furthermore, the model recorded 24 instances of False Negatives (FN), representing situations where positive cases were erroneously predicted as negative, indicating the occurrence of Type II errors.

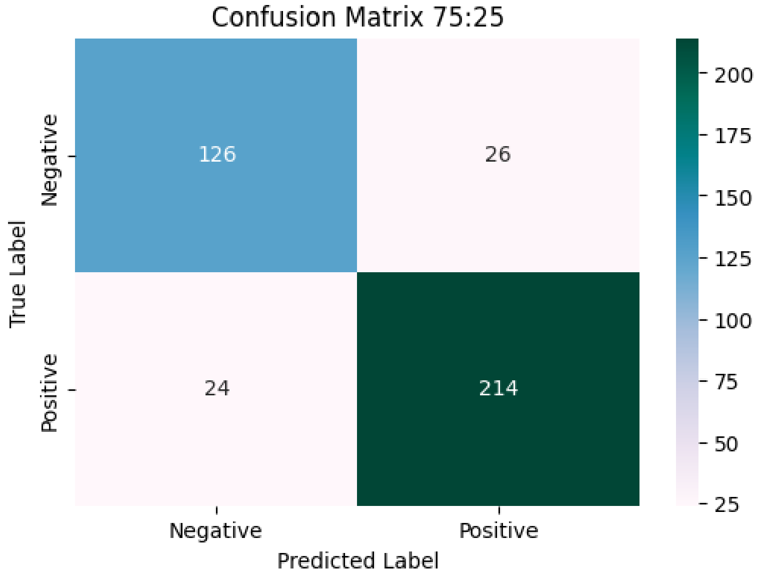

Figure 2. Confusion matrix data training and data testing 75:25

Based on figure 2's confusion matrix plot, confusion matrix yields compelling insights into the model's performance, showcasing 214 instances accurately identified as True Positives (TP), affirming the model's proficiency in correctly predicting positive cases. Concurrently, the model attains 126 instances of True Negatives (TN), effectively identifying negative cases. However, it also incurs 26 False Positives (FP), instances where the model erroneously classifies negative cases as positive, indicative of Type I errors. Additionally, the model registers 24 instances of False Negatives (FN), representing situations where positive cases are inaccurately predicted as negative, highlighting the occurrence of Type II errors.

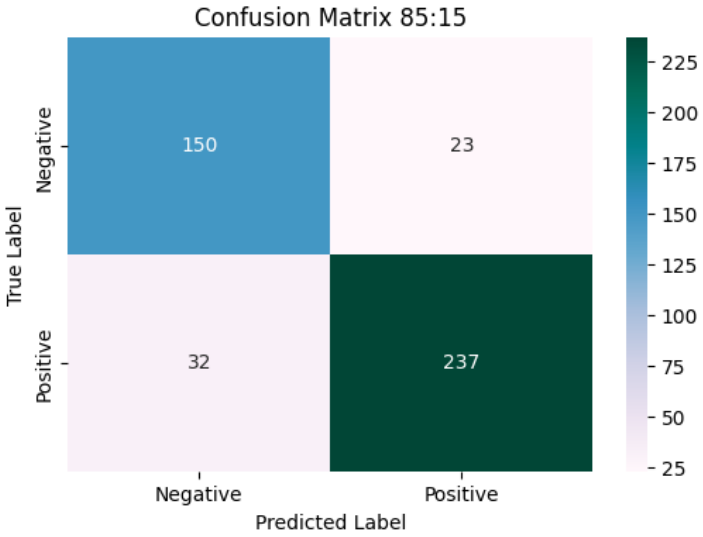

Figure 3. Confusion matrix data training and data testing 85:15

Based on confusion matrix plot in figure 3, examination of the confusion matrix reveals noteworthy metrics for the model’s performance 237 instances classified as True Positives (TP), signifying accurate identification of positive cases. Additionally, there are 150 instances of True Negatives (TN), reflecting the model’s proficiency in correctly recognizing negative cases. On the other hand, the model incurs 23 False Positives (FP), representing instances where it incorrectly identifies negative cases as positive. Furthermore, there are 32 instances of False Negatives (FN), indicating situations where positive cases are inaccurately predicted as negative.

Accuracy

Based on the results of the study, the accuracy results of each division of training and testing data can be seen in the table 2

Table 2. Accuracy of data training and data testing

No	split of data training and testing	Accuracy
1.	65:35	88.16%
2.	75:25	87.17%
3.	85:15	87.55%

From experimental results shown in table 2, it is known that the best accuracy results are obtained from the division of training and testing data of 65:35 with an accuracy value of 88.16%. The selection of a 65:35 training-to-testing data split, resulting in an accuracy of 88.16%, signifies a nuanced balance in the model development process. Allocating 65% of the data for training ensures a robust learning phase, enabling the model to discern intricate patterns and relationships within the dataset. This substantial training set minimizes the risk of overfitting, where the model might excessively tailor itself to the training data, including noise. Simultaneously, the 35% dedicated to testing serves as a comprehensive evaluation, challenging the model with a diverse range of instances to gauge its generalization capability effectively. The success of this split ratio may also stem from its alignment with the dataset's distribution and complexity, ensuring that both training and testing subsets are representative. The achieved high accuracy of 88.16% reflects the model's proficiency in capturing underlying patterns, providing a reliable indication of its performance on unseen data. It's important to note that the choice of an optimal split ratio often involves experimentation, considering the specific characteristics of the dataset for effective model training and evaluation.

Prescision

The insights into precision metrics are illuminated through table 3, offering a detailed breakdown and analysis of the obtained results.

Table 3. Prescision of data training and data testing

No	Split of data training and testing	Prescision
1.	65:35	0.883
2.	75:25	0.871
3.	85:15	0.877

From table 3 information, experimental findings underscore the optimal precision achieved when employing a data split of 65% for training and 35% for testing. In this configuration, the model demonstrates a superior ability to precisely identify relevant instances within the dataset, as evidenced by a precision value of 0.883. This indicates that, within the training set, the model excels in minimizing both false positives and false negatives, ensuring a higher proportion of positively identified instances are indeed relevant. The 65:35 division strikes a balance that allows the model to discern patterns and features effectively, leading to heightened precision in its predictions. This nuanced approach to data partitioning evidently contributes to the superior performance in precision metrics compared to other tested ratios.

Recall

Derived from the study outcomes, table 4 illustrates the recall outcomes for each split of training and testing data.

Table 4. Recall of data training and data testing

No	Split of data training and testing	Recall
1.	65:35	0.881
2.	75:25	0.871
3.	85:15	0.875

From experimental results shown in table 4, the optimal recall performance, as indicated by the experimental results, is achieved through the 65:35 split of training and testing data, resulting in a recall value of 0.881. This particular split ratio likely provides an effective balance between the amount of data used for training and testing, allowing the model to generalize well to new, unseen instances. The 65% allocated for training potentially ensures that the model captures the underlying patterns in the data, while the 35% reserved for testing ensures a robust evaluation on unseen samples. The higher recall value indicates that the model's improved capacity to accurately detect instances of interest may be attributed in part to this balancing.

F1-score

The comparative F1 score outcomes between the training and testing datasets are presented in table 5.

Table 5. F1-score of data training and data testing

No	Split of data training and testing	F1-score
1.	65:35	0.882
2.	75:25	0.871
3.	85:15	0.876

From table 5 information, the optimal performance, as indicated by the highest F1-score of 0.882, is achieved with a training and testing data split of 65:35. This result suggests that a larger portion of the dataset dedicated to training, specifically 65%, coupled with 35% for testing, yields the most favorable balance between precision and recall. A more extensive training dataset allows the model to learn the underlying patterns and intricacies of the data, enhancing its ability to generalize and make accurate predictions on unseen instances during testing. The 65:35 division strikes a pragmatic balance, preventing overfitting by providing sufficient diverse instances for testing while ensuring a robust model through comprehensive training.

CONCLUSION

In conclusion, after a thorough examination of diverse performance metrics, encompassing accuracy, precision, recall, and F1 score, the data split of 65:35 emerges as the clear frontrunner, outshining alternatives like 75:25 and 85:15 across all parameters. This supremacy is underscored by the remarkable results obtained from the 65:35 configuration, boasting an accuracy of 88.16%, precision at 0.883, recall registering 0.881, and a commendable F1 score of 0.882. The robustness of the 65:35 split lies in its adept balance between training and testing instances. Allocating 65% of the data to training enables the model to glean insights from a rich and diverse dataset, facilitating a nuanced understanding of underlying patterns. Simultaneously, the 35% assigned to testing ensures stringent validation, preventing overfitting and fortifying the model's generalizability. The cumulative effect of this balance maximizes overall model performance, reinforcing that the 65:35 data split is the most effective configuration for achieving optimal results in this experimental context.

ACKNOWLEDGEMENT

Author extends sincere gratitude to the Master's Program Department of Electrical Engineering, Universitas Islam Indonesia, for providing the environment and resources necessary for the successful completion of this study. Special thanks are due to Dr. Eng. Hendra Setiawan, S.T., M.T., for his invaluable contributions through insightful inputs and meaningful discussions.

REFERENCES

Abbas, M., Memon, K. A., Jamali, A. A., Memon, S., & Ahmed, A. (2019). Multinomial Naive Bayes classification model for sentiment analysis. *IJCSNS International Journal of Computer Science and Network Security*, *19*(3), 62-67.

Al-Mohaithef, M., Abdelmohsen, S. A., Algameel, M., & Abdelwahed, A. Y. (2022). Screening for identification of patients at high risk for diabetes-related foot ulcers: A cross-sectional study. *Journal of International Medical Research*, *50*(3), 03000605221087815. https://doi.org/10.1177/03000605221087815

Arokiasamy, P., Salvi, S., & Selvamani, Y. (2021a). *Global burden of diabetes mellitus: Prevalence, pattern, and trends*. In: Kickbusch, I., Ganten, D., Moeti, M. (eds) Handbook of Global Health. Springer, Cham.

https://doi.org/10.1007/978-3-030-05325-3_28-2

Association, A. D. (2019). Standards of medical care in diabetes—2019 abridged for primary care providers. *Clinical Diabetes: A Publication of the American Diabetes Association*, *37*(1), 11.

Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., & Sillero-Denamiel, M. R. (2021). Variable selection for Naïve Bayes classification. *Computers & Operations Research*, *135,* 105456. https://doi.org/10.1016/j.cor.2021.105456

Charley, E., Dinner, B., Pham, K., & Vyas, N. (2023). Diabetes as a consequence of acute pancreatitis. *World Journal of Gastroenterology*, *29*(31), 4736. https://doi.org/10.3748/Fwjg.v29.i31.4736

Chen, S., Webb, G. I., Liu, L., & Ma, X. (2020). A novel selective naïve Bayes algorithm. *Knowledge-Based Systems*, *192*, 105361. https://doi.org/10.1016/j.knosys.2019.105361

Esposito, S., Toni, G., Tascini, G., Santi, E., Berioli, M. G., & Principi, N. (2019). Environmental factors associated with type 1 diabetes. *Frontiers in Endocrinology*, *10,* 592. https://doi.org/10.3389/fendo.2019.00592

Hassan, M. M., Rony, M. A. T., Khan, M. A. R., Hassan, M. M., Yasmin, F., Nag, A., Zarin, T. H., Bairagi, A. K., Alshathri, S., & El-Shafai, W. (2023). Machine learning-based rainfall prediction: Unveiling insights and forecasting for improved preparedness. *IEEE Access*, *11*, 132196–132222. https://doi.org/10.1109/ACCESS.2023.3333876

Janež, A., Guja, C., Mitrakou, A., Lalic, N., Tankova, T., Czupryniak, L., Tabák, A. G., Prazny, M., Martinka, E., & Smircic-Duvnjak, L. (2020). Insulin therapy in adults with type 1 diabetes mellitus: A narrative review. *Diabetes Therapy*, *11,* 387–409. https://doi.org/10.6084/m9.figshare.11310668

Khursheed, R., Singh, S. K., Wadhwa, S., Kapoor, B., Gulati, M., Kumar, R., Ramanunny, A. K., Awasthi, A., & Dua, K. (2019). Treatment strategies against diabetes: Success so far and challenges ahead. *European Journal of Pharmacology*, *862,* 172625. https://doi.org/10.1016/j.ejphar.2019.172625

Kuo, F. Y., Cheng, K.-C., Li, Y., & Cheng, J.-T. (2021). Oral glucose tolerance test in diabetes, the old method revisited. *World Journal of Diabetes*, *12*(6), 786. https://doi.org/10.4239/wjd.v12.i6.786

Magkos, F., Hjorth, M. F., & Astrup, A. (2020). Diet and exercise in the prevention and treatment of type 2 diabetes mellitus. *Nature Reviews Endocrinology*, *16*(10), 545–555. https://doi.org/10.1038/s41574-020-0381-5

Maswadi, K., Ghani, N. A., Hamid, S., & Rasheed, M. B. (2021). Human activity classification using Decision Tree and Naive Bayes classifiers. *Multimedia Tools and Applications*, *80*, 21709–21726. https://doi.org/10.1007/s11042-020-10447-x

Patil, R., & Gothankar, J. (2019). Risk factors for type 2 diabetes mellitus: An urban perspective. *Indian Journal of Medical Sciences*, *71*(1), 16–21. https://doi.org/10.25259/IJMS_5_2019

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information*, *11*(4), 193. https://doi.org/10.3390/info11040193

Sousa, A. P., Cunha, D. M., Franco, C., Teixeira, C., Gojon, F., Baylina, P., & Fernandes, R. (2021). Which role plays 2-hydroxybutyric acid on insulin resistance? *Metabolites*, *11*(12), 835. https://doi.org/10.3390/metabo11120835

Wickramasinghe, I., & Kalutarage, H. (2021). Naive Bayes: Applications, variations and vulnerabilities: A review of literature with code snippets for implementation. *Soft Computing*, *25*(3), 2277–2293. https://doi.org/10.1007/s00500-020-05297-6