Algorithm Analysis Comparison of Naïve Bayes and Logistic Regression Methods for Predicting Diabetes

Windania Purba, Mardi Turnip, and Evta Indra

***Abstract*— Diabetes is caused by increased blood glucose (or blood sugar) levels in the body's metabolism, which causes severe damage to the heart, blood vessels, eyes, kidneys and nerves over time. Diabetes is a disease that cannot be cured but can be overcome by changing healthier lifestyle habits. The factor that causes the increase in diabetes is the delay in the early diagnosis of diabetes. Patients often die before receiving a diagnosis of diabetes due to complications from the condition. There are many types of variables and circumstances that can delay the diagnosis of diabetes. One of the efforts made in the early detection of diabetes is to utilize machine learning to assist in a fast and accurate diagnosis by modelling the calcifications of diabetes. This study uses a training set data analysis model by comparing the two modelling methods between Logistic Regression and Naïve Bayes. In this study, the Naïve Bayes algorithm got an accuracy of 89.5%. The ROC-AUC value is 0.895, and the precision value is 88 True Positive, 11 True Negative, 10 False Negative and 91 False Positive. Hence, it can be concluded that the precision for category 0 is 0.89 and category 1 is 0.90. Based on model making, both models have the same fit speed. Therefore, it can be concluded in the case of diabetes classification, the model that gets the best value is Logistic Regression.**

***Index Terms*— Diabetes, Naïve Bayes, Logistic regression, Data mining, Health information systems**.

I. INTRODUCTION

Diabetes is caused by increased blood glucose (or blood sugar) levels in the body's metabolism, which causes severe damage to the heart, blood vessels, eyes, kidneys, and nerves over time [1]. According to data released by [2], around 422 million people with diabetes globally live in low and middle-income countries, and diabetes is one of the deadliest diseases, with 1.5 million deaths each year. Over the last few decades, there has been a consistent increase in both the incidence and prevalence of diabetes.

When the blood supply to the heart stops completely, the unhealthy lifestyle is one of the causes of diabetes [3]. Not only parents but all ages can be affected by diabetes [4]. Diabetes is a disease that cannot be cured but can be overcome by changing healthier lifestyle habits [5][6]. This situation can be fatal to health in the future if not detected early [7].

Diabetes usually does not show clear enough symptoms, so the sufferer only finds out about it after damage to vital organs, heart, blood vessels, eyes, kidneys, and nerves [8].

W. Purba, M. Turnip, and E. Indra are with Faculty of Technology and Computer Science, University of Prima Indonesia, Medan, Indonesia (e-mails: windania@unprimdn.ac.id*, marditurnip@unprimdn.ac.id, evtapribadi@gmail.com).

The factor that causes an increase in diabetes is the delay in the early diagnosis of diabetes suffered by patients [9][10]. Patients often die before receiving a diagnosis of diabetes due to complications from the condition. There are many kinds of variables and circumstances that make the delay in diagnosing diabetes unnoticed [11]. If blood sugar levels are high or exceed normal values, making diabetes a disease that is quite dangerous [12][13]. According to sources [14] if the value of the glucose (blood sugar) level is less than 140 mg/dL, the level is considered normal, but if it is between 140 and 199 mg/dL, it means there is a prediabetes condition.

To prevent an increase in diabetes, early detection of patients is needed [15]. One of the efforts made in the early detection of diabetes is to utilize machine learning to assist in a fast and accurate diagnosis by modeling the calcifications of diabetes [16]. This classification model was created to determine whether a patient has diabetes. The model was built using training data set analysis by comparing the two modeling methods between Logistic Regression and Naïve Bayes. Future trends can be classified and predicted using a classification model [17].

Research related to the early detection of Diabetes has been carried out using various classification modeling methods, including those carried out by [18] by applying the Support Vector Machine (SVM) calcification method to get the highest accuracy value of 87%. Hence, it is necessary to improve with other classification methods, in this study a comparison will be made of 2 methods, namely Naïve Bayes [19] and Logistic Regression, to find the best prediction accuracy for Diabetes.

II. METHOD

A. Types of Research

This research was conducted to determine which algorithm has a higher accuracy value in the classification of diabetes diagnosis by comparing and evaluating the Naïve Bayes model with the Logistic Regression method. The method of collecting data for this research includes using information from the University of California, Irvine, a machine learning data repository, which can be accessed through the website [20].

A straightforward probability classifier, the Naive Bayes Classifier (NBC), can determine probability by adding up the frequency and value combinations in a given data set. The Nave Bayes approach is a statistical method for performing induction inference on a classification problem [21]. The benefit of employing Nave Bayes is that it only needs a small amount of training data (Training Data) to produce the necessary

parameter estimates for the classification process.

The second most famous machine learning algorithm is logistic regression. Logistic regression and linear regression are comparable in many ways. However, their utility is where the most significant difference lies [22]. While logistic regression is used for classification assignments, a linear regression approach is used to predict and estimate values.

B. *Work Procedures*

The research is planned by the flow diagram depicted in Fig. 1 so that it can proceed by the topics presented and be finished on schedule.

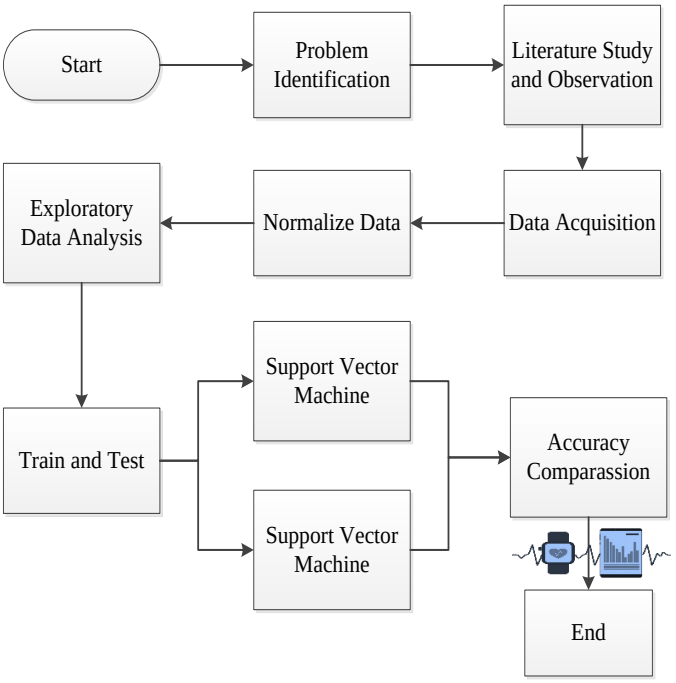


Fig. 1. Flow Diagram

This research followed a structured methodology consisting of several key stages. First, the problem was identified to establish the research focus. Next, data acquisition was conducted to gather relevant information. This was followed by a literature review and observations to build a strong theoretical foundation. The collected data was then normalized to ensure consistency and reliability. Subsequently, exploratory data analysis was performed to uncover patterns and insights. The core analysis involved comparing two predictive methods, such as Naïve Bayes and logistic regression to determine which provided the highest accuracy in predicting diabetes. Finally, a comprehensive comparison of the accuracy of both models was conducted to evaluate their effectiveness.

III. Result and Discussion

A. *Problem Analysis*

In this study, the Naïve Bayes and Logistic Regression algorithms will be used to classify data from diabetic patients so that the output can predict whether the patient has diabetes.

B. *Data Analysis*

In the dataset to be processed, there are 520 rows and 17 columns of data with 18 attributes which we can see in Table 1 below.

TABLE I
Dataset Attributes

Attributes
Age
Gender
Polyuria
Sudden Weight Loss
Weakness
Debilitation
Polyfagia
Genital Thrush
Visual Bluring
Itch
Delayed Healing
Partial Paresis
Muscle stiffness
Alopeciat
Obesity

C. *Data Analysis*

The first stage in data preparation is to remove duplicate data.

```
df=df.drop_duplicates()
df.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 251 entries, 0 to 519
Data columns (total 17 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Age                 251 non-null    int64
 1   Gender              251 non-null    object
 2   Polyuria            251 non-null    object
 3   Polydipsia          251 non-null    object
 4   sudden weight loss  251 non-null    object
 5   weakness            251 non-null    object
 6   Polyphagia          251 non-null    object
 7   Genital thrush      251 non-null    object
 8   visual blurring     251 non-null    object
 9   Itching             251 non-null    object
 10  Irritability        251 non-null    object
 11  delayed healing     251 non-null    object
 12  partial paresis     251 non-null    object
 13  muscle stiffness    251 non-null    object
 14  Alopecia            251 non-null    object
 15  Obesity             251 non-null    object
 16  class               251 non-null    object
dtypes: int64(1), object(16)
memory usage: 35.3+ KB
```

Fig. 2. Drop Duplicates

D. *Data Analysis*

Data normalization converts parameter values from string format to integer to facilitate machine learning modeling. The following table shows the normalization procedure.

TABLE II
DATASET

```
# PENGAMBILAN DATA DARI GOOGLE DRIVE
df=pd.read_csv(r'/content/drive/MyDrive/KlasifikasiDiabetes/diabetes_data_upload (2).csv')
df.head()
```

1 to 5 of 5 entries Filter

Index	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	Male	No	Yes	No	Yes	No	No	No	Yes	No	Yes	No	Yes	Yes	Yes	Positive
1	58	Male	No	No	No	Yes	No	No	Yes	No	No	No	Yes	No	Yes	No	Positive
2	41	Male	Yes	No	No	Yes	Yes	No	No	Yes	No	Yes	No	Yes	Yes	No	Positive
3	45	Male	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	No	Positive
4	60	Male	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Positive

TABLE III
DATASET NORMALIZATION

Dataset Normalization

```
[ ] df.replace({'Negative': 0, 'Positive': 1}, inplace=True)
    df.replace({'No': 0, 'Yes': 1}, inplace=True)

[ ] df.head()
```

	Age	Gender	Polyuria	Polydipsia	sudden weight loss	weakness	Polyphagia	Genital thrush	visual blurring	Itching	Irritability	delayed healing	partial paresis	muscle stiffness	Alopecia	Obesity	class
0	40	Male	0	1	0	1	0	0	0	1	0	1	0	1	1	1	1
1	58	Male	0	0	0	1	0	0	1	0	0	0	1	0	1	0	1
2	41	Male	1	0	0	1	1	0	0	1	0	1	0	1	1	0	1
3	45	Male	0	0	1	1	1	1	0	1	0	1	0	0	0	0	1
4	60	Male	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1

E. Balancing Dataset

The processed dataset has unbalanced data due to a large number of negative data from positive data so it is necessary to process the data balancing between positive data and negative data. We can see the data balancing process in Fig. 3:

Dataset Balancing

```
[ ] print(df['class'].value_counts())
    cls_0=df[df['class']==0]
    cls_1=df[df['class']==1]

    1    173
    0     78
    Name: class, dtype: int64

[ ] cls_0=cls_0.sample(500,replace=True)
    cls_1=cls_1.sample(500,replace=True)
    df=pd.concat([cls_0,cls_1],axis=0)
    df.info()

    <class 'pandas.core.frame.DataFrame'>
    Int64Index: 1000 entries, 207 to 76
    Data columns (total 17 columns):
     #   Column              Non-Null Count  Dtype
    ---  ------              --------------  -----
     0   Age                 1000 non-null   int64
     1   Gender              1000 non-null   object
     2   Polyuria            1000 non-null   int64
     3   Polydipsia          1000 non-null   int64
     4   sudden weight loss  1000 non-null   int64
     5   weakness            1000 non-null   int64
     6   Polyphagia          1000 non-null   int64
     7   Genital thrush      1000 non-null   int64
     8   visual blurring     1000 non-null   int64
     9   Itching             1000 non-null   int64
     10  Irritability        1000 non-null   int64
     11  delayed healing     1000 non-null   int64
     12  partial paresis     1000 non-null   int64
     13  muscle stiffness    1000 non-null   int64
     14  Alopecia            1000 non-null   int64
     15  Obesity             1000 non-null   int64
     16  class               1000 non-null   int64
    dtypes: int64(16), object(1)
    memory usage: 140.6+ KB
```

Fig. 3. Dataset Balancing

F. Balancing Dataset

1. Correlation Heatmaps

It can be concluded in Fig. 4 that the correlation between the parameters/features in the dataset is not related, or all parameters/features are essential in determining the pattern.

2. Diabetes Group By Sex

Figure 5 shows that the distribution of diabetes by gender is about 700 for men and about 300 for women.

3. Diabetes Group By Sex

After creating a visualization of diabetes based on gender, the next task is to visualize diabetes based on age groups in terms of gender. The plot can be seen in the Fig. 6.

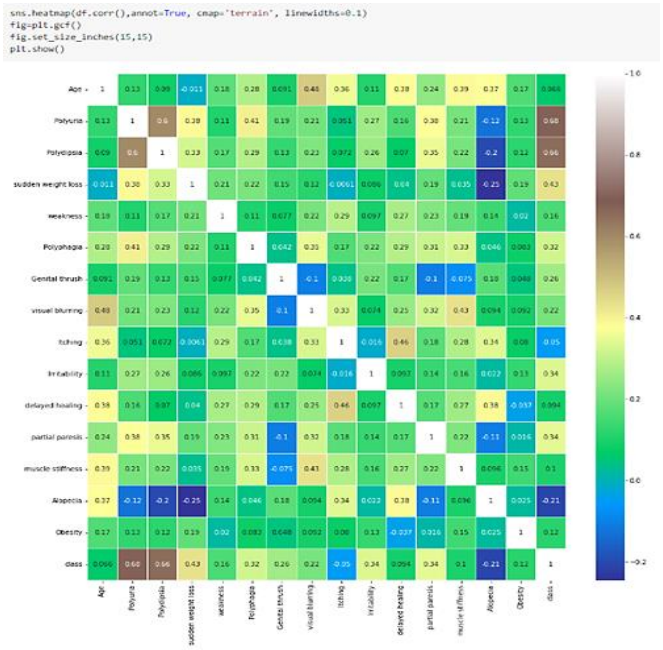


Fig. 4. Correlation Heatmap

```
by_gender = df.groupby('Gender')
by_gender.size().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3f773e41d0>
```

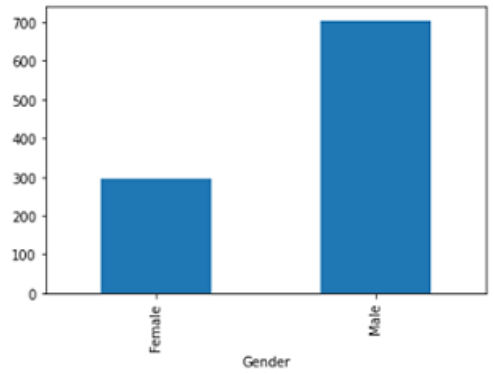


Fig. 5. Group by Gender

```
by_gender_age = df.groupby(['category','Gender'])
by_gender_age.size().plot(kind='bar')
plt.rcParams["figure.figsize"] = (5,5)
plt.show()
```

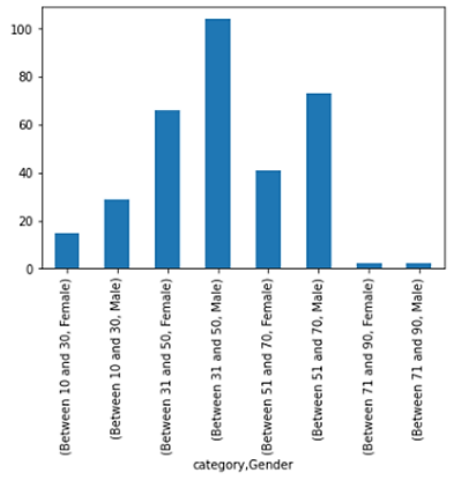


Fig. 6. Group by Gender and Sex

G. *Balancing Dataset*

At the data partition stage, the dataset will be split into two parts: the training and testing datasets. Partition configuration will be carried out with 80 percent training data and 20 percent testing data (80:20). which we can see in the figure below:

```
[ ] X = df.drop(['Gender','ids_counted','category','class'],axis=1)

[ ] Y= df['class']

[ ] from sklearn.model_selection import train_test_split
    from sklearn.metrics import accuracy_score

    X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size= 0.2, random_state=101,shuffle=True)

[ ] Y_train.value_counts()

    0    401
    1    399
    Name: class, dtype: int64

[ ] Y_test.value_counts()

    1    101
    0     99
    Name: class, dtype: int64
```

Fig. 7. Dataset Splitting

H. *Data Training and Test*

Data training will be conducted using two basic classification algorithms: Logistic Regression and naive Bayes. In this study, the Logistic Regression algorithm was created with the default configuration, and the Naïve Bayes algorithm was created with the default configuration.

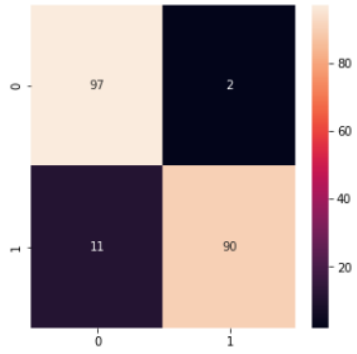
1. *Logistic Regression*

In this study, the Logistic Regression algorithm got an accuracy of 93.5%. The ROC-AUC value is 0.935, and the precision value is 97 True Positive, 2 True Negative, 11 False Negative, and 90 False Positive Thus it can be concluded that the precision in category 0 is 0.98, while in category 1 it is 0.89. which we can see in the figure below:

```
lc_matrix = confusion_matrix(Y_test, y_pred_lc)

sns.heatmap(lc_matrix, annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc06d17a810>
```

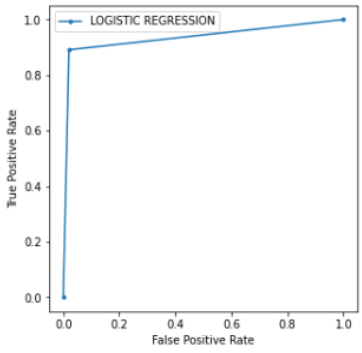



```
lc_auc = roc_auc_score(Y_test, y_pred_lc)
print('LOGISTIC REGRESSION: ROC AUC=%.3f' % (lc_auc))

lc_fpr, lc_tpr, _ = roc_curve(Y_test, y_pred_lc)

plt.plot(lc_fpr, lc_tpr, marker='.', label='LOGISTIC REGRESSION')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()
plt.show()
```

```
LOGISTIC REGRESSION: ROC AUC=0.935
```




```
print(classification_report(y_pred_lc, Y_test))
```

	precision	recall	f1-score	support
0	0.98	0.90	0.94	108
1	0.89	0.98	0.93	92
accuracy			0.94	200
macro avg	0.94	0.94	0.93	200
weighted avg	0.94	0.94	0.94	200

Fig. 8. Accuracy Logistic Regression

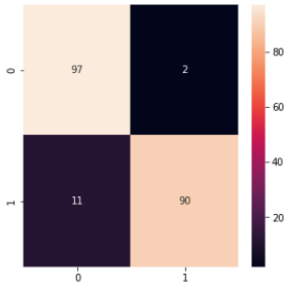
2. *Naïve Bayes*

In this study, the Naïve Bayes algorithm got an accuracy of 89.5%. The ROC-AUC value is 0.895, and the precision value is 88 True Positive, 11 True Negative, 10 False Negative, and 91 False Positive. Thus, it can be concluded that the precision for category 0 is 0.89, while for category 1, it is 0.90.

```
lc_matrix = confusion_matrix(Y_test, y_pred_lc)

sns.heatmap(lc_matrix, annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fc06d17a810>
```

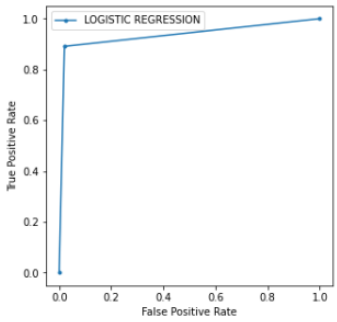



```
lc_auc = roc_auc_score(Y_test, y_pred_lc)
print('LOGISTIC REGRESSION: ROC AUC=%.3f' % (lc_auc))

lc_fpr, lc_tpr, _ = roc_curve(Y_test, y_pred_lc)

plt.plot(lc_fpr, lc_tpr, marker='.', label='LOGISTIC REGRESSION')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.legend()
plt.show()
```

```
LOGISTIC REGRESSION: ROC AUC=0.935
```




```
print(classification_report(y_pred_lc, Y_test))
```

```
              precision    recall  f1-score   support

           0       0.98      0.90      0.94       108
           1       0.89      0.98      0.93        92

    accuracy                           0.94       200
   macro avg       0.94      0.94      0.93       200
weighted avg       0.94      0.94      0.94       200
```

Fig. 9. Accuracy Naïve bayes

I. Comparisson Method

With the results of making the algorithm above, Logistic Regression has the highest accuracy value, the highest ROC-AUC value, and the highest precision value. Based on the speed of model creation, both Logistic Regression and Naïve Bayes models have the same fit speed. Thus, it can be concluded that in the case of diabetes classification, the best model is Logistic Regression.

```
print("Accuracy Logistic Regression",'\n',classification_report(y_pred_lc, Y_test),'\n')
print("Accuracy Naive Bayes",'\n',classification_report(y_pred_nb, Y_test))
```

```
Accuracy Logistic Regression
              precision    recall  f1-score   support

           0       0.98      0.90      0.94       108
           1       0.89      0.98      0.93        92

    accuracy                           0.94       200
   macro avg       0.94      0.94      0.93       200
weighted avg       0.94      0.94      0.94       200


Accuracy Naive Bayes
              precision    recall  f1-score   support

           0       0.89      0.90      0.89        98
           1       0.90      0.89      0.90       102

    accuracy                           0.90       200
   macro avg       0.89      0.90      0.89       200
weighted avg       0.90      0.90      0.90       200
```

Fig. 10. Comparassion Accuracy

IV. Conclusion

To prevent an increase in diabetes, early detection of patients is needed. One of the efforts made in the early detection of diabetes is to utilize machine learning to assist in a fast and accurate diagnosis by modeling the calcifications of diabetes. With the results of making the algorithm above, Logistic Regression has the highest accuracy value, the highest ROC-AUC value, and the highest precision value. Based on the speed of model creation, both Logistic Regression and Naïve Bayes models have the same fit speed. Thus, it can be concluded that for diabetes classification, the best model to use is Logistic Regression.

Declaration of Competing Interest

This research is fully supported by the Faculty of Science and Technology, Universitas Prima Indonesia, Medan, Indonesia. Deepest appreciation to Universitas Prima Indonesia, and also to all friends (senior and junior) who participated in the research.

Acknowledgement

This research is fully supported by Prima Indonesia University Information System study program. We would like to thank our esteemed lecturers at Prima Indonesia University for their continuous guidance, help, and input.

References

[1] N. G. Ramadhan, "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus," *Scientific Journal of Informatics*, vol. 8, no. 2, 2021, doi: 10.15294/sji.v8i2.32484

[2] "Diabetes." https://www.who.int/health-topics/diabetes#tab=tab_1 (accessed Nov. 22, 2022).

[3] The Lancet Diabetes & Endocrinology, "Diabetes behind bars: challenging inadequate care in prisons," *Lancet Diabetes Endocrinol.*, vol. 6, no. 5, p. 347, 2018, doi: 10.1016/S2213-8587(18)30103-7.

[4] A. Hussain and A. J. M. Boulton, "COVID-19 and diabetes: International diabetes federation perspectives," *Diabetes Res. Clin. Pract.*, vol. 167, p. 108339, 2020, doi: 10.1016/j.diabres.2020.108339.

[5] C. Irace, "Awareness of Diabetes Complication in Subjects with Type 2 Diabetes," *Diabetes Obes. Int. J.*, vol. 7, no. 1, pp. 1–5, 2022, doi: 10.23880/doij-16000251.

[6] Y. Granillo and G. H. Goldsztein, "Machine Learning as a Tool to the Diagnosis of Diabetes," *Journal of Student Research*, vol. 11, no. 1, 2022, doi: 10.47611/jsrhs.v11i1.2513

[7] D. Sitanggang, E. Indra, J. H. Gulo, and M. Turnip, "Implementation of the K-Nearest Neighbor Algorithm for Detecting Heart Attack Disease," *Internetworking Indonesia Journal*, vol. 13, no. 2, pp. 35-41, 2021.

[8] M. Ranjit Reddy, P. Lakshmi Sagar, and N. S. Shaik, "Diabetes Mellitius Detection and Self Management based on Machine Learning," *J Pharm NegatResults*, vol. 13, no. 4, 2022, doi: 10.47750/pnr.2022.13.04.138

[9] K. M. Kaka-Khan, H. Mahmud, and A. A. Ali, "Rough Set-Based Feature Selection for Predicting Diabetes Using Logistic Regression with Stochastic Gradient Decent Algorithm," *UHD Journal of Science and Technology*, vol. 6, no. 2, 2022, doi: 10.21928/uhdjst. v6n2y2022.pp85-93

[10] A. F. N. Masruriyah, H. Y. Novita, and C. E. Sukmawati, "Performance Evaluation of Popular Supervised Learning Algorithms Towards Cardiovascular Disease," *Computer Science (CO-SCIENCE)*, vol. 8, no. 3, pp. 420–426, 2023, doi: 10.32493/informatika. v8i3.34103.

[11] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods," *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/3820360

[12] Y. Murai, K. Miyajima, M. Shinohara, T. Yamada, and T. Ohta, "Investigation of Pharmacological Responses to an Anti-Diabetic Drug Pioglitazone in Female Spontaneously Diabetic Torii (SDT) Fatty Rats, A New Obese Type 2 Diabetic Rat," *Clin. Diabetes Res.*, vol. 1, no. 1, pp. 8–13, 2017, doi: 10.36959/647/488

[13] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, "Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms," *Neural Comput Appl*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.

[14] A. B. Amjoud and M. Amrouch, "Transfer Learning for Automatic Image Orientation Detection Using Deep Learning and Logistic Regression," *IEEE Access*, vol. 10, pp. 128543–128553, 2022, doi: 10.1109/ACCESS.2022.3225455.

[15] T. M. Le, T. M. Vo, T. N. Pham, and S. V. T. Dao, "A Novel Wrapper-Based Feature Selection for Early Diabetes Prediction Enhanced with a Metaheuristic," *IEEE Access*, vol. 9, pp. 7869–7884, 2021, doi: 10.1109/ACCESS.2020.3047942

[16] C. K. Suryadevara, "Issue 4 Diabetes Risk Assessment Using Machine Learning: A Comparative Study of Classification Algorithms," 2023. [Online]. Available: www.iejrd.com

[17] L. Shrinivasan, R. Verma, and M.D. Nandeesh, "Early prediction of diabetes diagnosis using hybrid classification techniques," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 3, pp. 1139–1148, 2023. https://doi.org/10.11591/ijai.v12.i3.pp1139-1148

[18] F. Fersellia, E. Utami, and A.Yaqin, "Sentiment Analysis of Shopee Food Application User Satisfaction Using the C4.5 Decision Tree Method," *Sinkron*, vol. 8, no. 3, pp. 1554–1563, 2023. https://doi.org/10.33395/sinkron. v8i3.12531

[19] K. A. Rahman, B. Irawan, C. Setianingsih, I. Hadi, R. E. Putra, M. H. Haekal, and M. I. Junando, "Crime Rate Mapping in Bandung City Area with Online Media Data by Using Naïve Bayes Method*," Internetworking Indonesia Journal*, vol. 14, no. 1, pp. 21-26, 2022.

[20] "Index of /ml/machine-learning-databases/00529," https://archive.ics.uci .edu/ml/machine-learning-databases/00529/ (accessed Nov. 23, 2022).

[21] C. Hong Sheng, T. So Ha, and K. Khalid Abdul, "Therapeutic Agents Targeting at AGE-RAGE Axis for the Treatment of Diabetes and Cardiovascular Disease: A Review of Clinical Evidence," *Clin. Diabetes Res.*, vol. 1, no. 1, pp. 16–34, 2017, doi: 10.36959/647/490.

[22] A. Oana, N. Jane, B. Richie, W. Peter, and M. Richard WA, "Disposition Index (DI) is not Improved with High-Intensity Intermittent Exercise in Adults with Hyperinsulinemia and Pre-Diabetes," *Clin. Diabetes Res.*, vol. 5, no. 1, pp. 55–61, 2021, doi: 10.36959/647/495.