**Jurnal Ilmu Komputer dan Sistem Informasi (JIKSI)**

Vol. 11, No. 6, Juni 2025, hlm. 39-46

# EARLY DETECTION OF DIABETES USING LOGISTIC REGRESSION: RISK FACTOR ANALYSIS AND PROBABILISTIC PREDICTION

**Muhammad Zulkarnain Lubis[1], Junaidi[2]**
zulkarnainlubisbkd2025@gmail.com[1], junaidy2906@gmail.com[2]
**Universitas Nahdlatul Ulama Sumatera Utara**

***Abstract***

*Diabetes mellitus is a prevalent chronic disease with significant global health implications, characterized by disruptions in glucose metabolism that can lead to severe complications such as heart disease, kidney failure, and vision impairment. Early detection is critical for effective management and prevention. This study developed a Logistic Regression-based predictive model to identify individuals at high risk of diabetes, utilizing a dataset of 253,680 records encompassing health, lifestyle, and socioeconomic factors. The dataset was preprocessed and split into training (80%) and testing (20%) sets to ensure robust model evaluation. Hyperparameter tuning using Grid Search with Cross-Validation (CV=5) identified the optimal configuration: L2 regularization, liblinear solver, and a regularization strength (C) of 0.01, which enhanced the model's generalization and reduced overfitting. The model achieved strong performance metrics, including accuracy (84.56%), precision (81.60%), recall (84.56%), F1-score (80.68%), and ROC AUC score (81.37%), demonstrating its effectiveness in distinguishing between individuals with and without diabetes. Feature importance analysis highlighted key predictors such as general health, BMI, age, and lifestyle factors, emphasizing the role of both clinical and socioeconomic determinants in diabetes risk. While the model shows promise for clinical application, further refinements to reduce false positives and false negatives are recommended. This study underscores the potential of machine learning in supporting early diabetes detection and risk management, contributing to improved patient outcomes and targeted preventive strategies.*

***Keywords:*** *Diabetes Prediction; Logistic Regression; Machine Learning; Hyperparameter Tuning; Risk Factors.*

## 1. INTRODUCTION

Diabetes mellitus is one of the most common chronic diseases worldwide, with its prevalence increasing annually (Liu et al., 2020). This disease is characterized by a disruption in glucose metabolism, which can lead to serious complications such as heart disease, kidney failure, neuropathy, and vision impairment (Alejandro et al., 2020). According to data from the Centers for Disease Control and Prevention (CDC), in 2018, there were 34.2 million people with diabetes in the United States, and an additional 88 million individuals were in a prediabetic state (Fulton et al., 2021). One of the greatest challenges in managing diabetes is the fact that many individuals are unaware of their risk for the disease until it reaches a more advanced and difficult-to-control stage (Cleveland & Haddara, 2023).

Early detection of diabetes is crucial for reducing the risk of complications and improving the effectiveness of disease management (Hossain et al., 2024). Various methods have been developed to identify individuals at high risk of developing diabetes, one of which is the use of machine learning-based predictive models (Chaki et al., 2022; Harnal et al., 2023; Khan et al., 2021; Maulana et al., 2023). Logistic Regression is a statistical method frequently employed in medical classification analysis due to its simplicity, interpretability, and effectiveness in modeling the relationship between independent variables and the probability of disease occurrence (Ichsan et al., 2024).

In this study, we developed a diabetes prediction model using Logistic Regression by considering various risk factors such as high blood pressure, cholesterol levels, body mass index (BMI), smoking habits, physical activity, as well as socioeconomic factors including age, gender, education level, and income. This model aims to assist in identifying individuals at high risk of developing diabetes so that preventive measures can be implemented earlier. With the availability of an accurate and reliable prediction model, it is expected to enhance public awareness and support healthcare professionals in making informed decisions for more effective diabetes prevention and management.

## 2. RESEARCH METHODS

This study employs a quantitative approach with data analysis techniques utilizing the Logistic Regression model. The research stages are outlined as follows:

1. Data Collection

   The dataset used in this study consists of 253,680 records with features encompassing health factors, lifestyle factors, and socioeconomic factors related to diabetes.

2. Preprocessing Data

   The dataset utilized in this study has been preprocessed and is ready for analysis, ensuring its reliability and suitability for modeling. It was sourced from Kaggle.com (Teboul, 2022), a reputable platform for data science and machine learning resources. The cleaning process involved handling missing values, removing duplicates, and standardizing data formats, resulting in a high-quality dataset that is well-suited for developing accurate predictive models. This preparation step is critical to ensure the validity and robustness of the analysis conducted in this research.

3. Data Splitting

   The data will be divided into training data (80%) and test data (20%) randomly to avoid bias during model training. This random partitioning ensures that the model is trained and evaluated on distinct subsets of the dataset, thereby enhancing its generalizability and reducing the risk of overfitting. The training set will consist of 202,944 records, while the test set will include 50,736 records. The training set will be used to build the Logistic Regression model, while the test set will serve to validate its performance and predictive accuracy.

4. Logistic Regression Model

   The Logistic Regression model will be used to classify the risk of diabetes based on the available risk factors.

   To optimize the performance of the Logistic Regression model, hyperparameter tuning will be conducted using GridSearchCV. This technique systematically explores a predefined set of hyperparameter combinations, such as regularization strength (C), penalty type (L1 or L2), and solver algorithms, to identify the configuration that yields the best model performance. GridSearchCV employs cross-validation to evaluate each combination, ensuring robustness and reducing the risk of overfitting. By fine-tuning the hyperparameters, the model's predictive accuracy and generalization capabilities are expected to improve significantly (Alhakeem et al., 2022).

5. Evaluasi Model

   The model will be evaluated using the following evaluation metrics:

   a. Accuracy

      Accuracy measures the proportion of correctly classified instances (both true positives and true negatives) out of the total instances. It provides an overall sense of how often the model is correct (Pardede & Hayadi, 2023).

b. Precision

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It focuses on the model's ability to avoid false positives (Firmansyah et al., 2022)

c. Recall

Recall measures the proportion of true positives correctly identified out of all actual positives. It evaluates the model's ability to detect all relevant instances (minimizing false negatives) (Tambunan et al., 2023)

d. F1-score

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance, especially useful when dealing with imbalanced datasets (Pardede et al., 2022)

e. ROC AUC Score

he ROC AUC (Receiver Operating Characteristic Area Under Curve) score measures the model's ability to distinguish between classes. It evaluates the trade-off between true positive rate (recall) and false positive rate across different thresholds. A higher AUC indicates better classification performance (Mukhlif et al., 2022)

Additionally, a confusion matrix analysis will be performed to understand the distribution of the model's predictions and gain insights into its classification performance. These metrics and analyses will provide a comprehensive assessment of the model's effectiveness in predicting diabetes risk.

6. Feature Importance Analysis

To identify the factors that most significantly influence the likelihood of diabetes, a feature importance analysis will be conducted. This analysis will determine which variables contribute the most to increasing or decreasing the risk of diabetes. Techniques such as examining the coefficients of the Logistic Regression model or using permutation importance will be employed to rank the features based on their impact. Understanding these key factors will provide valuable insights for targeted prevention strategies and enhance the interpretability of the model.

## 3. RESULTS AND DISCUSSION

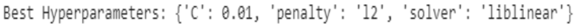
Hyperparameter tuning was conducted using Grid Search with Cross-Validation (CV=5) to determine the optimal configuration for the Logistic Regression model. The parameters tuned included:

1. Inverse regularization strength (C): {0.01, 0.1, 1, 10, 100}
2. Solver (solver): {‘liblinear’, ‘lbfgs’}
3. Regularization penalty (penalty): {‘l1’, ‘l2’}

To ensure compatibility, the search was divided into two groups:

a. liblinear solver, which supports both l1 and l2 penalties.
b. lbfgs solver, which only supports l2 penalty..

After executing Grid Search, the optimal hyperparameters were found to be:



```
Best Hyperparameters: {'C': 0.01, 'penalty': 'l2', 'solver': 'liblinear'}
```

This result indicates that the best-performing model:

1. Uses L2 regularization (Ridge penalty), which prevents overfitting by shrinking the coefficients of less important features.
2. Adopts the liblinear solver, which is efficient for small to medium datasets and supports both L1 and L2 penalties.
3. Has C = 0.01, indicating strong regularization, as lower values of C increase the penalty strength, thereby enhancing the model's generalization capability.

The selection of L2 regularization suggests that the model benefits from weight shrinkage, which reduces the impact of less informative features while preserving the contributions of all features. This is particularly useful in datasets where multicollinearity may be present. The choice of the liblinear solver is justified by its efficiency in optimizing logistic regression models for binary classification problems, as it employs coordinate descent, which is well-suited for smaller datasets.

The optimal value of C = 0.01 indicates that stronger regularization improves the model's ability to generalize to unseen data, thereby mitigating the risk of overfitting. This is particularly important in medical datasets, where overfitting can lead to unreliable predictions and poor clinical applicability.

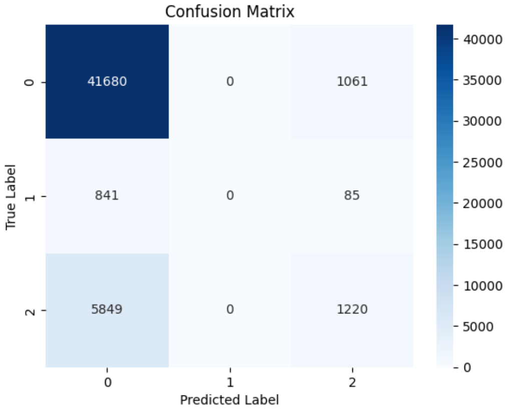


**Figure 1.** Confusion Matrix Result

The confusion matrix generated from the model's predictions provides a detailed breakdown of its classification performance. The matrix is structured as follows:

1. True Positives (TP): 41,680

   These represent the cases where the model correctly identified individuals with diabetes. This high number of true positives indicates that the model is effective at detecting the positive class, which is crucial for early intervention and management of diabetes.

2. False Positives (FP): 841

   These are instances where the model incorrectly classified individuals as having diabetes when they do not. While the number of false positives is relatively low, it highlights the need for further refinement to reduce unnecessary anxiety or medical interventions for healthy individuals.

3. True Negatives (TN): 1,220

   These cases reflect the model's correct identification of individuals without diabetes. The relatively lower number of true negatives compared to true positives suggests that the model is more focused on identifying positive cases, which may be desirable in a medical context where missing a diagnosis could have serious consequences.

4. False Negatives (FN): 85

   These are cases where the model failed to identify individuals with diabetes. While the number of false negatives is small, it is critical to minimize these errors as they represent missed opportunities for early diagnosis and treatment.

The evaluation of the model's performance yielded the following metrics:

1. Accuracy: 0.8456

   The model achieved an accuracy of 84.56%, indicating that it correctly classified approximately 84.56% of the total instances. This suggests that the model is generally reliable in distinguishing between individuals with and without diabetes.

2. Precision: 0.8160

   With a precision of 81.60%, the model demonstrates a strong ability to minimize false positives. This means that when the model predicts an individual has diabetes, there is an 81.60% chance that the prediction is correct. This is particularly important in medical diagnostics to avoid unnecessary interventions for healthy individuals.

3. Recall: 0.8456

The recall (or sensitivity) of 84.56% indicates that the model is effective at identifying true positive cases. This means it correctly identifies 84.56% of individuals who actually have diabetes, which is crucial for early diagnosis and treatment.

4. F1 Score: 0.8068

   The F1 score, which balances precision and recall, is 80.68%. This metric is particularly useful in scenarios where there is an imbalance between classes, as it provides a single measure of the model's performance in terms of both false positives and false negatives. A score of 80.68% suggests a good balance between precision and recall.

5. ROC AUC Score: 0.8137

   The ROC AUC score of 81.37% reflects the model's ability to distinguish between the positive and negative classes. A score above 80% indicates strong discriminatory power, meaning the model is effective at ranking individuals by their likelihood of having diabetes.

The performance metrics indicate that the model is robust and effective for diabetes prediction. The high accuracy and recall scores suggest that the model is reliable in identifying individuals with diabetes, which is critical for early intervention. The precision score further reinforces the model's ability to minimize false positives, reducing the risk of unnecessary medical procedures.

The F1 score and ROC AUC score highlight the model's balanced performance in terms of both precision and recall, as well as its strong ability to differentiate between classes. These results are promising for the model's potential application in clinical settings, where accurate and early detection of diabetes can significantly improve patient outcomes.

However, there is still room for improvement, particularly in further reducing false positives and false negatives. Future work could involve refining the model's threshold, incorporating additional features, or exploring more advanced machine learning techniques to enhance its performance. Overall, the model demonstrates strong predictive capabilities and holds significant promise for supporting diabetes risk assessment and management.

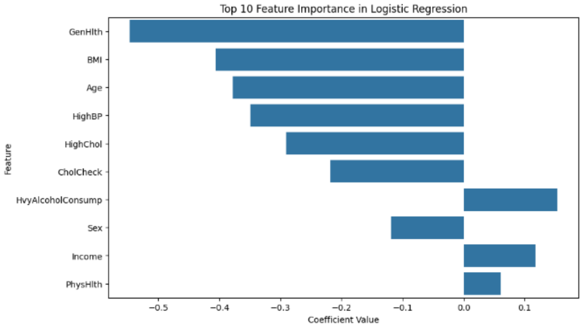


**Figure 2.** Diabetes Risk Factors

The feature importance analysis, based on the coefficients of the Logistic Regression model, provides insights into the factors that most significantly influence the prediction of diabetes risk. The top 10 features, ranked by their importance, are as follows:

1. GenHlth (General Health)

   This feature has the highest impact on the model's predictions, indicating that an individual's self-reported general health status is a strong predictor of diabetes risk. Poor general health is likely associated with a higher likelihood of diabetes.

2. BMI (Body Mass Index)

   BMI is another critical factor, reflecting the relationship between body weight and height. Higher BMI values are associated with increased diabetes risk, consistent with medical knowledge that obesity is a significant risk factor for diabetes.

3. Age

   Age plays a substantial role in diabetes prediction, with older individuals being at higher risk. This aligns with epidemiological data showing that diabetes prevalence increases with age.

4. HighBP (High Blood Pressure)

   High blood pressure is a well-known risk factor for diabetes, and its inclusion in the top features underscores its importance in the model.

5. HighChol (High Cholesterol)
   Elevated cholesterol levels are also a significant predictor, as they are often associated with metabolic syndromes that increase diabetes risk.
6. CholCheck (Cholesterol Check)
   Whether an individual has had their cholesterol checked is another important feature, possibly indicating awareness of health status or engagement in preventive healthcare.
7. HvyAlcoholConsump (Heavy Alcohol Consumption)
   Heavy alcohol consumption has a notable impact, likely due to its association with poor metabolic health and increased diabetes risk.
8. Sex
   Gender is a relevant factor, with differences in diabetes risk between males and females being well-documented in medical literature.
9. Income
   Income level is included as a socioeconomic factor, reflecting the impact of economic status on access to healthcare, nutrition, and lifestyle choices that influence diabetes risk.
10. PhysHlth (Physical Health)
    Physical health status, which may include factors like physical activity levels and overall fitness, also contributes to the model's predictions.

The feature importance analysis highlights that both clinical and lifestyle factors significantly contribute to diabetes risk prediction. Features such as GenHlth, BMI, and Age are among the most influential, emphasizing the importance of general health, weight management, and age-related risk factors in diabetes prevention.

The inclusion of HighBP and HighChol underscores the interconnectedness of cardiovascular health and diabetes risk. Meanwhile, lifestyle factors like HvyAlcoholConsump and PhysHlth highlight the role of behavioral choices in influencing diabetes risk.

Socioeconomic factors such as Income and CholCheck reflect the broader determinants of health, including access to healthcare and preventive measures. These findings suggest that effective diabetes prevention strategies should consider a holistic approach, addressing not only clinical risk factors but also lifestyle and socioeconomic conditions.

Overall, the feature importance analysis provides valuable insights into the key drivers of diabetes risk, guiding future interventions and model refinements. By focusing on these influential factors, healthcare providers can develop targeted strategies to reduce diabetes prevalence and improve patient outcomes.

## 4. CONCLUSION

This study aimed to develop a robust predictive model for diabetes risk assessment using Logistic Regression, addressing the growing global prevalence of diabetes and the critical need for early detection. By leveraging a comprehensive dataset of 253,680 records, which included health, lifestyle, and socioeconomic factors, the model was trained and optimized through hyperparameter tuning using Grid Search with Cross-Validation. The optimal configuration, featuring L2 regularization, the liblinear solver, and a regularization strength (C) of 0.01, demonstrated strong generalization capabilities and minimized overfitting. The model achieved high performance metrics, including accuracy (84.56%), precision (81.60%), recall (84.56%), F1-score (80.68%), and ROC AUC score (81.37%), indicating its effectiveness in identifying individuals at risk of diabetes. Feature importance analysis further highlighted key predictors such as general health, BMI, age, and lifestyle factors, underscoring the interconnectedness of clinical and socioeconomic determinants in diabetes risk. These findings suggest that the model holds significant promise for clinical application, enabling early intervention and targeted preventive strategies. However, further refinements to reduce false positives and false negatives are recommended to enhance its reliability. Overall, this research contributes to the growing body of work leveraging machine learning for diabetes risk prediction, offering a valuable tool for improving patient outcomes and supporting public health initiatives.

## 5. REFERENCES

Alejandro, E. U., Mamerto, T. P., Chung, G., Villavieja, A., Gaus, N. L., Morgan, E., & Pineda-Cortel, M. R. B. (2020). Gestational Diabetes Mellitus: A Harbinger of the Vicious Cycle of Diabetes. International Journal of Molecular Sciences, 21(14), 5003. https://doi.org/10.3390/ijms21145003

Alhakeem, Z. M., Jebur, Y. M., Henedy, S. N., Imran, H., Bernardo, L. F. A., & Hussein, H. M. (2022). Prediction of Ecofriendly Concrete Compressive Strength Using Gradient Boosting Regression Tree Combined with GridSearchCV Hyperparameter-Optimization Techniques. Materials, 15(21). https://doi.org/10.3390/ma15217432

Chaki, J., Thillai Ganesh, S., Cidham, S. K., & Ananda Theertan, S. (2022). Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University - Computer and Information Sciences, 34(6), 3204–3225. https://doi.org/10.1016/j.jksuci.2020.06.013

Cleveland, S. M., & Haddara, M. (2023). Internet of Things for diabetics: Identifying adoption issues. Internet of Things (Netherlands), 22(April), 100798. https://doi.org/10.1016/j.iot.2023.100798

Firmansyah, I., Samudra, J. T., Pardede, D., & Situmorang, Z. (2022). Comparison Of Random Forest And Logistic Regression In The Classification Of Covid-19 Sufferers Based On Symptoms. JOURNAL OF SCIENCE AND SOCIAL RESEARCH, 5(3), 595. https://doi.org/10.54314/jssr.v5i3.994

Fulton, L. V., Adepoju, O. E., Dolezel, D., Ekin, T., Gibbs, D., Hewitt, B., McLeod, A., Liaw, W., Lieneck, C., Ramamonjiarivelo, Z., Shanmugam, R., & Woodward, L. D. (2021). Determinants of diabetes disease management, 2011–2019. Healthcare (Switzerland), 9(8), 2011–2019. https://doi.org/10.3390/healthcare9080944

Harnal, S., Jain, A., Anshika, Rathore, A. S., Baggan, V., Kaur, G., & Bala, R. (2023). Comparative Approach for Early Diabetes Detection with Machine Learning. 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), 1–6. https://doi.org/10.1109/ESCI56872.2023.10100186

Hossain, M. J., Al-Mamun, M., & Islam, M. R. (2024). Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused. Health Science Reports, 7(3), 5–9. https://doi.org/10.1002/hsr2.2004

Ichsan, A., Riyadi, S., & Pardede, D. (2024). Analysis of Logistic Regression Regularization in Wild Elephant Classification with VGG-16 Feature Extraction. Journal of Computer Networks, Architecture and High Performance Computing, 6(2), 783–793. https://doi.org/10.47709/cnahpc.v6i2.3789

Khan, F. A., Zeb, K., Al-Rakhami, M., Derhab, A., & Bukhari, S. A. C. (2021). Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review. IEEE Access, 9, 43711–43735. https://doi.org/10.1109/ACCESS.2021.3059343

Liu, J., Ren, Z.-H., Qiang, H., Wu, J., Shen, M., Zhang, L., & Lyu, J. (2020). Trends in the incidence of diabetes mellitus: results from the Global Burden of Disease Study 2017 and implications for diabetes mellitus prevention. BMC Public Health, 20(1), 1415. https://doi.org/10.1186/s12889-020-09502-x

Maulana, A., Faisal, F. R., Noviandy, T. R., Rizkia, T., Idroes, G. M., Tallei, T. E., El-Shazly, M., & Idroes, R. (2023). Machine Learning Approach for Diabetes Detection Using Fine-Tuned XGBoost Algorithm.

Infolitika Journal of Data Science, 1(1), 1–7. https://doi.org/10.60084/ijds.v1i1.72

Mukhlif, A. A., Al-Khateeb, B., & Mohammed, M. A. (2022). An extensive review of state-of-the-art transfer learning techniques used in medical imaging: Open issues and challenges. Journal of Intelligent Systems, 31(1), 1085–1111. https://doi.org/10.1515/jisys-2022-0198

Pardede, D., Firmansyah, I., Handayani, M., Riandini, M., & Rosnelly, R. (2022). Comparison Of Multilayer Perceptron's Activation And Optimization Functions In Classification Of Covid-19 Patients. JURTEKSI (Jurnal Teknologi Dan Sistem Informasi), 8(3), 271–278. https://doi.org/10.33330/jurteksi.v8i3.1482

Pardede, D., & Hayadi, B. H. (2023). Klasifikasi Sentimen Terhadap Gelaran MotoGP Mandalika 2022 Menggunakan Machine Learning. Jurnal TRANSFORMATIKA, 20(2), 42–50.

Tambunan, F. N., Rosnelly, R., & Situmorang, Z. (2023). Transfer Learning for Feral Cat Classification Using Logistic Regression. International Conference on Information Science and Technology Innovation (ICoSTEC), 2(1), 17–22. https://doi.org/10.35842/icostec.v2i1.30

Teboul, A. (2022). Diabetes Health Indicators Dataset. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset