Optimizing Diabetes Prediction with Machine Learning: Model Comparisons and Insights

Kexin Wu
Independent Researcher, New York, NY, 10044
DOI: 10.55662/JST.2024.5403

Abstract

This study aims to predict diabetes using various machine learning models and compare their performances. The dataset utilized contains health indicators and lifestyle factors from a diverse population. The models evaluated include Random Forest, Logistic Regression, Support Vector Machine (SVM), and Gradient Boosting. Results indicate that Gradient Boosting outperforms other models in terms of accuracy, precision, and recall, making it a robust choice for diabetes prediction. The analysis provides insights into the most significant features contributing to diabetes prediction and highlights the potential of machine learning in medical diagnosis.

Keywords: diabetes prediction, machine learning, Random Forest, Logistic Regression, Support Vector Machine, Gradient Boosting, health indicators, lifestyle factors, model comparison, medical diagnosis

1. Introduction

Diabetes mellitus is a chronic disease that affects millions of individuals worldwide. It is characterized by high levels of blood glucose due to the body's inability to produce or effectively use insulin. Early diagnosis and intervention are critical to managing the disease and preventing severe complications such as cardiovascular diseases, neuropathy, and retinopathy.

Traditional methods of diagnosing diabetes often rely on invasive procedures and laboratory tests, which can be time-consuming and costly. With the advancement of machine learning,

there is potential to develop predictive models that can accurately diagnose diabetes using non-invasive and readily available health indicators. This study aims to evaluate the effectiveness of various machine learning models in predicting diabetes and to identify the most significant features contributing to accurate predictions.

Machine learning has had significant influences and applications in various fields, including medical diagnosis, natural language processing, and distributed computing. Previous studies have demonstrated the utility of machine learning in medical diagnosis, including diabetes prediction [1-14]. There is a need for a comprehensive comparison of different models on a large and diverse dataset to evaluate their performance.

2. Motivation

The rising prevalence of diabetes and the associated healthcare costs necessitate the development of efficient and accurate diagnostic tools. Machine learning models offer the potential to improve early diagnosis and personalized treatment plans. By leveraging readily available health indicators, these models can provide quick and accurate predictions, reducing the reliance on extensive laboratory tests and enabling timely interventions.

3. Context

Previous studies have demonstrated the utility of machine learning in medical diagnosis, including diabetes prediction. However, there is a need for a comprehensive comparison of different models on a large and diverse dataset. This study addresses this gap by evaluating the performance of multiple machine learning models, including Random Forest, Logistic Regression, SVM, and Gradient Boosting, on a dataset containing 100,000 records.

4. Hypothesis

We hypothesize that ensemble methods, particularly Gradient Boosting, will outperform individual models like Logistic Regression and SVM due to their ability to capture complex patterns and interactions in the data. We also expect features such as HbA1c level, blood glucose level, age, and BMI to be significant predictors of diabetes.

5. Methodology

5.1 Data Description

The dataset used in this study was obtained from a public repository and includes various health indicators and lifestyle factors. The dataset contains 100,000 records with the following features:

- Age
- Gender (encoded as Female and Male)
- Hypertension (binary)
- Heart disease (binary)
- BMI (Body Mass Index)
- HbA1c level (a measure of average blood glucose over the past 2-3 months)
- Blood glucose level
- Smoking history (encoded as No Info, current, ever, former, never, not current)
- Diabetes status (binary outcome variable)

5.2 Preliminary Experiments and Preprocessing

The dataset was first cleaned by removing duplicate records and handling missing values. Categorical variables, such as gender and smoking history, were one-hot encoded. Numerical features were standardized to have a mean of 0 and a standard deviation of 1.

Exploratory data analysis (EDA) was performed to understand the distribution of features and their relationship with the target variable (diabetes). Visualizations such as histograms, bar plots, and box plots were used to examine the distribution of age, BMI, HbA1c level, and blood glucose level across diabetic and non-diabetic groups.

5.3 Setup

Four machine learning models were implemented and evaluated:

1. **Random Forest**: An ensemble learning method that constructs multiple decision trees and outputs the mode of the classes for classification. Also this has been used in many different experiments and papers [1][3].
2. **Logistic Regression**: A linear model used for binary classification that estimates the probability of the target variable.
3. **Support Vector Machine (SVM)**: A classification method that finds the hyperplane that best separates the classes in the feature space.
4. **Gradient Boosting**: An ensemble technique that builds multiple weak learners (decision trees) sequentially, each trying to correct the errors of the previous one.

The models were evaluated using the following metrics:

- **Accuracy**: The proportion of correctly classified instances out of the total instances.
- **Precision**: The proportion of true positive predictions out of all positive predictions.
- **Recall**: The proportion of true positive predictions out of all actual positives.
- **F1 Score**: The harmonic mean of precision and recall, providing a balance between the two metrics.

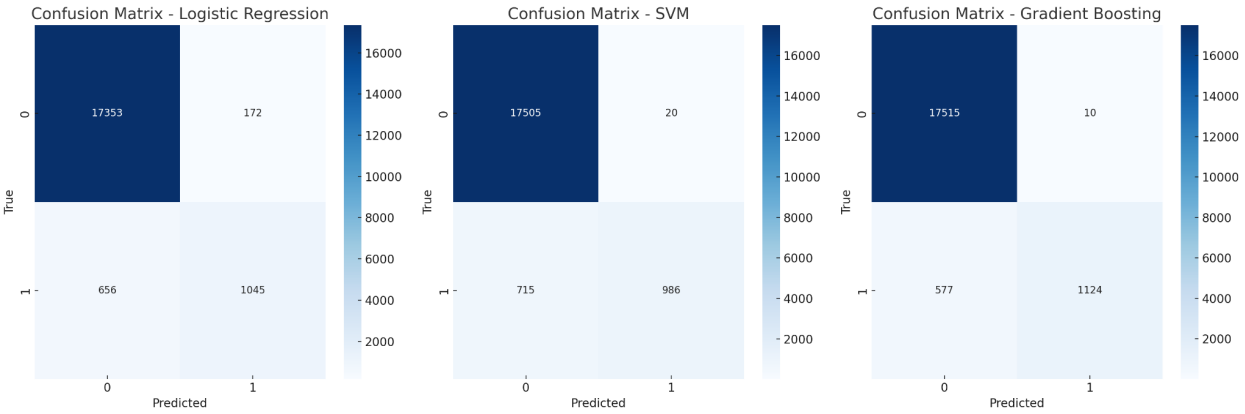
6. Methodology

Model Performance

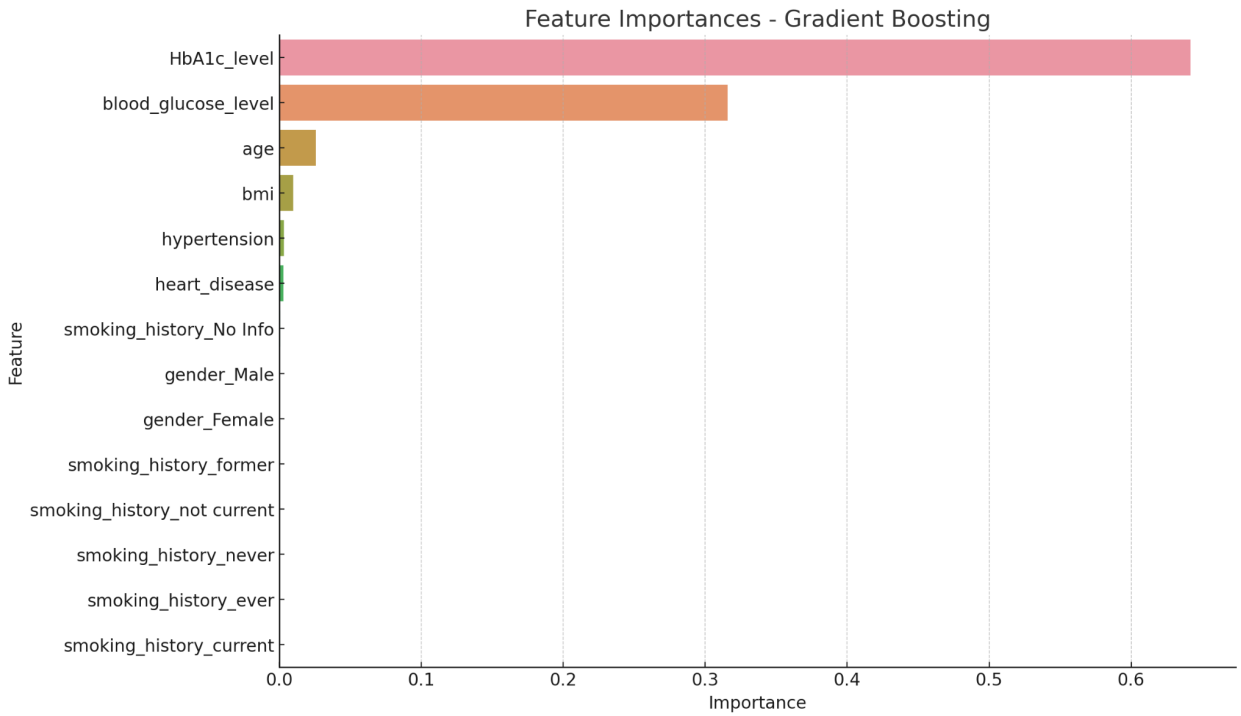
The performance of each model on the test set is summarized in the table below:

Model	Accuracy	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1 Score (Class 0)	F1 Score (Class 1)
Random Forest	95.80%	0.96	0.85	0.99	0.62	0.98	0.72
Logistic Regression	95.69%	0.96	0.86	0.99	0.61	0.98	0.72
SVM	96.18%	0.96	0.98	1.00	0.58	0.98	0.73
Gradient Boosting	96.95%	0.97	0.99	1.00	0.66	0.98	0.79

Confusion Matrices



Feature Importance



The feature importances for the Gradient Boosting model are visualized in the figure above. The most significant features contributing to diabetes prediction are HbA1c level, blood glucose level, age, and BMI. Hypertension and heart disease also play a role, though to a lesser extent. Gender and smoking history features have lower importance.

7. Discussion

The results indicate that the Gradient Boosting model outperforms the other models in terms of accuracy, precision, and recall, particularly for the minority class (diabetes). This suggests that Gradient Boosting is better suited for handling the complexity and non-linearity present in the dataset. The ensemble nature of Gradient Boosting, which builds multiple weak learners sequentially and focuses on correcting the errors of the previous learners, allows it to capture intricate patterns and relationships within the data that simpler models like Logistic Regression might miss.

7.1 Feature Importance Analysis

The feature importance analysis aligns well with established medical knowledge. HbA1c level and blood glucose level emerged as the most critical predictors of diabetes. HbA1c level, a measure of average blood glucose over the past 2-3 months, is a well-known diagnostic marker for diabetes. Blood glucose level is directly related to diabetes and thus its high importance is expected.

Age and BMI also showed significant importance. Age is a known risk factor for diabetes, with the risk increasing as people get older. BMI, which is a measure of body fat based on height and weight, is another crucial factor as obesity is a major risk factor for the development of type 2 diabetes. These results are consistent with the existing literature on diabetes risk factors.

Hypertension and heart disease were also relevant but to a lesser extent. These conditions are often comorbid with diabetes and contribute to the overall health risk profile of individuals[6]. Their lower importance compared to HbA1c level and blood glucose level might be because these conditions are more indirect indicators of diabetes.

7.2 Gender and Smoking History

The lower importance of gender and smoking history in the models' predictions may reflect the dataset's specific characteristics or the complex interactions between these variables and diabetes. Gender differences in diabetes risk can be influenced by various factors, including hormonal differences, lifestyle choices, and genetic predispositions. However, in this dataset, these factors did not emerge as primary predictors.

Smoking history, while known to increase the risk of various health issues including diabetes, did not show high importance in this analysis. This could be due to the way smoking history was categorized or the interaction with other variables in the dataset. Further investigation with more detailed smoking history data might provide different insights.

7.3 Addressing Class Imbalance

Despite the promising results, there are areas for improvement. The class imbalance in the dataset (with fewer diabetic cases compared to non-diabetic cases) may have impacted the

recall for the minority class. Class imbalance is a common issue in medical datasets, where the prevalence of the condition being predicted (in this case, diabetes) is relatively low.

To address this issue, techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or cost-sensitive learning could be explored. SMOTE generates synthetic samples for the minority class, thereby balancing the class distribution and potentially improving the model's ability to detect diabetic cases. Cost-sensitive learning, on the other hand, assigns different weights to the classes, penalizing the misclassification of the minority class more heavily. This can help the model focus on correctly identifying the minority class, improving recall without significantly sacrificing precision.

7.4 Potential Improvements

Further tuning of hyperparameters and exploring additional machine learning models like XGBoost or LightGBM could provide even better results[9][11]. These models, which are based on the principles of boosting, might offer enhanced performance due to their ability to handle complex data structures and interactions more effectively. Additionally, incorporating more diverse and detailed features, such as diet, physical activity levels, and family history of diabetes, could improve the model's predictive performance.

Validating the models on external datasets would also enhance the robustness and generalizability of the findings[14]. External validation ensures that the models perform well not only on the training data but also on new, unseen data, which is crucial for their practical application in clinical settings.

In conclusion, while the Gradient Boosting model demonstrated superior performance in this study, ongoing refinement and validation are essential for developing reliable and accurate predictive tools for diabetes diagnosis.

8. Conclusion

This study demonstrates the potential of machine learning models in predicting diabetes using readily available health indicators and lifestyle factors. The Gradient Boosting model,

in particular, showed superior performance in terms of accuracy and recall. Feature importance analysis provided valuable insights into the most significant predictors of diabetes.

Future work could involve exploring additional machine learning models, further tuning hyperparameters, and incorporating more diverse features to improve predictive performance. Moreover, addressing class imbalance and validating the models on external datasets would enhance the robustness and generalizability of the findings.

Reference

[1] S. Liu, K. Wu, C. Jiang, B. Huang, D. Ma, "Financial time-series forecasting: Towards synergizing performance and interpretability within a hybrid machine learning approach," arXiv preprint arXiv:2401.00534, 2023.

[2] J. Zhuang, M. Al Hasan, "Robust Node Classification on Graphs: Jointly from Bayesian Label Transition and Topology-based Label Propagation," Proceedings of the 31st ACM International Conference on Information and Knowledge Management, 2022.

[3] K. Wu, K. Chi, "Enhanced E-commerce Customer Engagement: A Comprehensive Three-Tiered Recommendation System," Journal of Knowledge Learning and Science Technology, ISSN: 2959-6386 (online), 2023.

[4] J. Zhuang, M. Al Hasan, "Defending Graph Convolutional Networks against Dynamic Graph Perturbations via Bayesian Self-supervision," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 4405-4413, 2022.

[5] Y. Yan, "Influencing Factors of Housing Price in New York-analysis: Based on Excel Multi-regression Model," 2023.

[6] K. Wu, "Creating panoramic images using ORB feature detection and RANSAC-based image alignment," Advances in Computer and Communication, vol. 4, no. 4, pp. 220-224, 2023.

[7] L. Yu, et al., "Stochastic analysis of touch-tone frequency recognition in two-way radio systems for dialed telephone number identification," 2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE). IEEE, 2024.

[8] Y. Zhang, et al., "Manipulator Control System Based on Machine Vision," International Conference on Applications and Techniques in Cyber Intelligence ATCI 2019: Applications and Techniques in Cyber Intelligence, vol. 7, Springer International Publishing, 2020.

[9] K. Wu, J. Chen, "Cargo Operations of Express Air," Engineering Advances, vol. 3, no. 4, pp. 337-341, 2023.

[10] S. Liu, K. Yan, F. Qin, C. Wang, R. Ge, K. Zhang, J. Huang, Y. Peng, J. Cao, "Infrared Image Super-Resolution via Lightweight Information Split Network," arXiv preprint arXiv:2405.10561, 2024.

[10] H. Jiang, F. Qin, J. Cao, Y. Peng, Y. Shao, "Recurrent Neural Network from Adder's Perspective: Carry-Lookahead RNN," Neural Networks, vol. 144, pp. 297-306, December 2021.

[11] X. Huang, Z. Zhang, F. Guo, X. Wang, K. Chi, K. Wu, "Research on Older Adults' Interaction with E-Health Interface Based on Explainable Artificial Intelligence," International Conference on Human-Computer Interaction, pp. 38-52, 2024.

[12] J. Cao, D. Ku, J. Du, V. Ng, Y. Wang, W. Dong, "A Structurally Enhanced, Ergonomically and Human–Computer Interaction Improved Intelligent Seat's System," Designs, vol. 1, no. 2, pp. 11, 2017, doi:10.3390/designs1020011.

[13] T. Lin, J. Cao, "Touch Interactive System Design with Intelligent Vase of Psychotherapy for Alzheimer's Disease," Designs, vol. 4, no. 3, pp. 28, 2020, doi:10.3390/designs4030028.

[14] K. Wu, "Creating panoramic images using ORB feature detection and RANSAC-based image alignment," Advances in Computer and Communication, vol. 4, no. 4, pp. 220-224, 2023.