Improving Diabetes Prediction Accuracy Using Ensemble Machine Learning Models

Aadit Singh

ABSTRACT

This study investigates prediction of HbA1c level which is a principal biomarker of diabetes control based on patient biographical and health data from a publicly accessible dataset [1]. I tried regression models like Linear Regression [2], Decision Tree Regressor [3], and Random Forest Regressor [4] to predict accurate HbA1c levels. Upon facing poorly performing models, most likely because of data bias and feature insufficiency, I restructured the task as a classification problem by approximating the ranges of HbA1c levels into significant categories. I implemented models including Random Forest Classifier [5], Decision Tree Classifier [6], K-Nearest Neighbors [7], and an ensemble Voting Classifier [8]. The Voting Classifier increased the best accuracy to 72.5%, improving over Random Forest's standalone accuracy of 68.1% [5]. Model tuning focused on parameters such as the number of trees and maximum depth. Variance Inflation Factor analysis was executed to evaluate feature multicollinearity and it confirmed that multicollinearity was not a major issue. Results show that classification models are more suitable for this dataset and confirm the importance of feature engineering and hyperparameter adjustment. This finding demonstrates that classification models better suit this dataset, showing how predictive instruments can assist medical personnel in approximating HbA1c values without resorting to decisions purely based on costly or time-consuming laboratory testing.

KEYWORDS

HbA1c; diabetes prediction; machine learning; Random Forest; Voting Classifier; Kaggle; classification; glycemic control; ensemble learning; predictive modeling

INTRODUCTION

Diabetes has been present in my life for years as my mother has had it for years, and watching her daily struggles with blood sugar control brought me to consider how technology could improve her situation. Through this project I explore whether machine learning algorithms can be used to accurately predict HbA1c levels, which is a key indicator of long term glucose regulation, based on basic health and demographic data. Being able to accurately predict these levels might mean earlier intervention and improved outcomes for someone like my mom. While prior research has shown potential for the application of data science to medical prediction, complications such as small datasets and imbalanced features persist. Moreover, the prospect of forecasting HbA1c levels without depending on invasive laboratory measures could make early identification more straightforward in resource-poor or high volume clinical settings, improving results at scale. This project combines personal interest with data driven inquiry in evaluating the performance of different algorithms and testing their potential utility for supporting diabetes management.

LITERATURE REVIEW

The use of machine learning for the prediction of glycated hemoglobin (HbA1c) has gained popularity in recent years due to its promise of improved early diagnosis and control of diabetes. Researchers have explored various methods of predicting HbA1c using clinical, demographic, and lifestyle information, often trying to create predictive models that could potentially eliminate invasive or costly laboratory testing.

In the study "Improving Current Glycated Hemoglobin Prediction in Adults: Use of Machine Learning Algorithms With Electronic Health Records" [9] published in JMIR Medical Informatics (2021), Alhassan et al. investigated the use of a range of machine learning models—such as logistic regression, random forests, and multilayer perceptrons—to predict if adult patient HbA1c levels were above the threshold (≥5.7%) using electronic health record (EHR) data. Their models both incorporated longitudinal and historical patient data. The top performing model was a multilayer perceptron (MLP) that, using longitudinal data, achieved 83.2% accuracy. Age and random blood sugar were the most important features in their work. While the paper effectively showed how EHR-based longitudinal data can be used to improve prediction, it concentrated on binary classification (normal or elevated HbA1c) and utilized high quality clinical histories. My work differs in that it operates with a much smaller and readily available feature set, for example, cholesterol, triglycerides, and simple demographics and frames HbA1c prediction as a three-class classification problem (Normal, Prediabetic, Diabetic). This approach focuses on broader availability in real world, resource constrained environments where longitudinal information may not be easily accessible.

Another notable study, "Predicting Three-Month Fasting Blood Glucose and Glycated Hemoglobin Using Ensemble Learning" [10], was published in Scientific Reports in 2023. A large dataset of over 375,000 Chinese type 2 diabetic patients was used to forecast subsequent HbA1c and fasting blood glucose. The authors built an ensemble of machine learning algorithms, including a specially designed random forest, to forecast whether patients would meet the criterion of HbA1c control (<7%) at a three month follow up. Their best performing models had excellent performance with area under the curve (AUC) values up to 0.97. Variables to be predicted were BMI, baseline blood glucose, adherence to medication, and diet, each of which required repeat follow ups and close clinical monitoring. By contrast, my study includes one time snapshot rather than follow up over time of patient information, and uses standard medical thresholds (5.7 and 6.5 HbA1c) to define categories. Further, while my project does not simply report AUC or binary values, it compares multiple classifiers—namely, Decision Tree, K-Nearest Neighbors, and Random Forest—before merging them into a Voting Classifier, improving classification performance to 72.5%.

Both experiments illustrate the power of machine learning in medical prediction issues but rely on large volume, high detail datasets readily available only in large clinical systems. My own research attempts to discern whether such valuable predictions can be realized with less sophisticated, more widely available inputs. It examines how reformulating the problem from one of regression to one of classification may lower model complexity and enhance interpretability. In addition, the emphasis on ensemble techniques such as the Voting Classifier illustrates how relatively simple algorithms might be combined for better performance when specifically put

together. This way, my work attempts to bridge the gap between highly sophisticated clinical modeling and more practical aids to frontline triage and screening in diabetes.

METHODS

This study used a publicly available diabetes dataset from Kaggle [1], consisting of 1,000 patients with demographic and clinical data such as age, gender, urea, cholesterol, triglycerides, HDL, creatinine ratio and HbA1c levels. I aimed to forecast the level of HbA1c, which is a main indicator of blood sugar, using this data. These features were chosen because they are typically linked with insulin resistance and metabolic health, therefore being potentially useful markers of glucose regulation in the long term (HbA1c). For instance, elevated triglycerides and reduced HDL are likely to be present in combination with poor glycemic control, and measurements of urea and creatinine can reflect diabetic complications like renal impairment.

Preprocessing of data included data imputation and data cleaning. Missing data were handled by treating zero values in measures and using median values to replace them. Normalization of data was achieved by utilizing standard procedures to make data suitable for feature magnitude-based algorithms such as K-Nearest Neighbors [7]. For multicollinearity checks I computed Variance Inflation Factors for all variables and they were all less than 10 which assured that multicollinearity was not a major issue. I initially attempted to forecast HbA1c values using regression models like Linear Regression [2], Decision Tree Regressor [3], and Random Forest Regressor [4].

I used Linear Regression [2] as the default model to find out whether the input features were significantly linearly correlated with the HbA1c levels. I chose the Decision Tree Regressor to find out if there was any non-linear trend or decision based thresholds in the data. Finally, I used a Random Forest Regressor [4], an ensemble of numerous decision trees, to avoid overfitting and combine model predictions for superior overall predictability. R-squared (R^2), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were used to evaluate the performance of models. As I achieved low values of R^2 and high errors (See Figure 1) I converted the task to a classification task by binning HbA1c to three bins (See Figure 2).

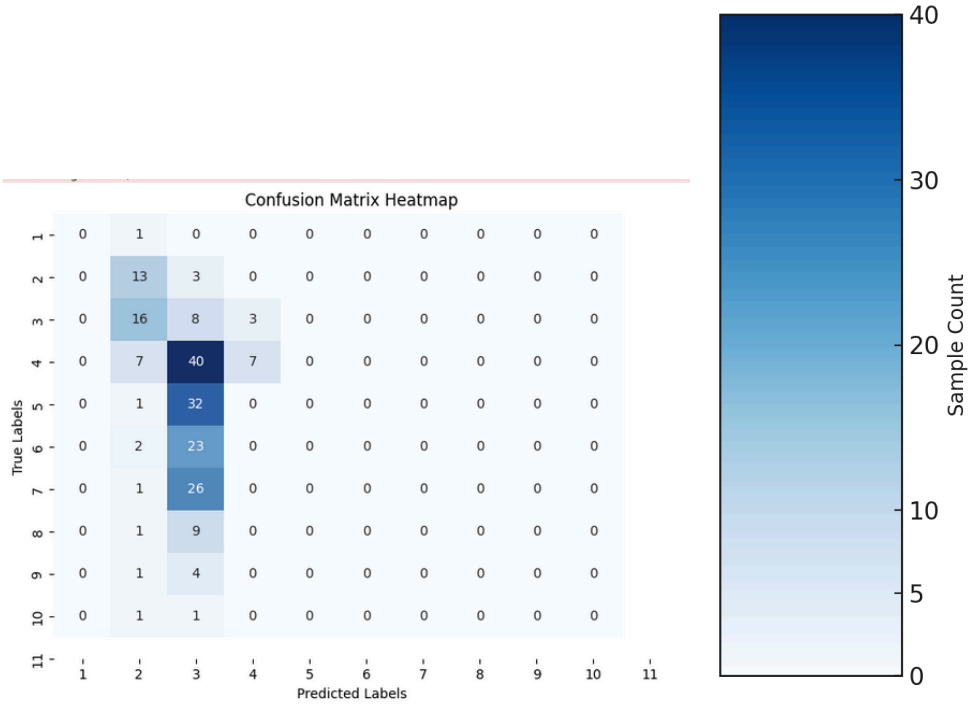

Fig. 1 Confusion Matrix Heatmap (Before shortening to 3 bins)

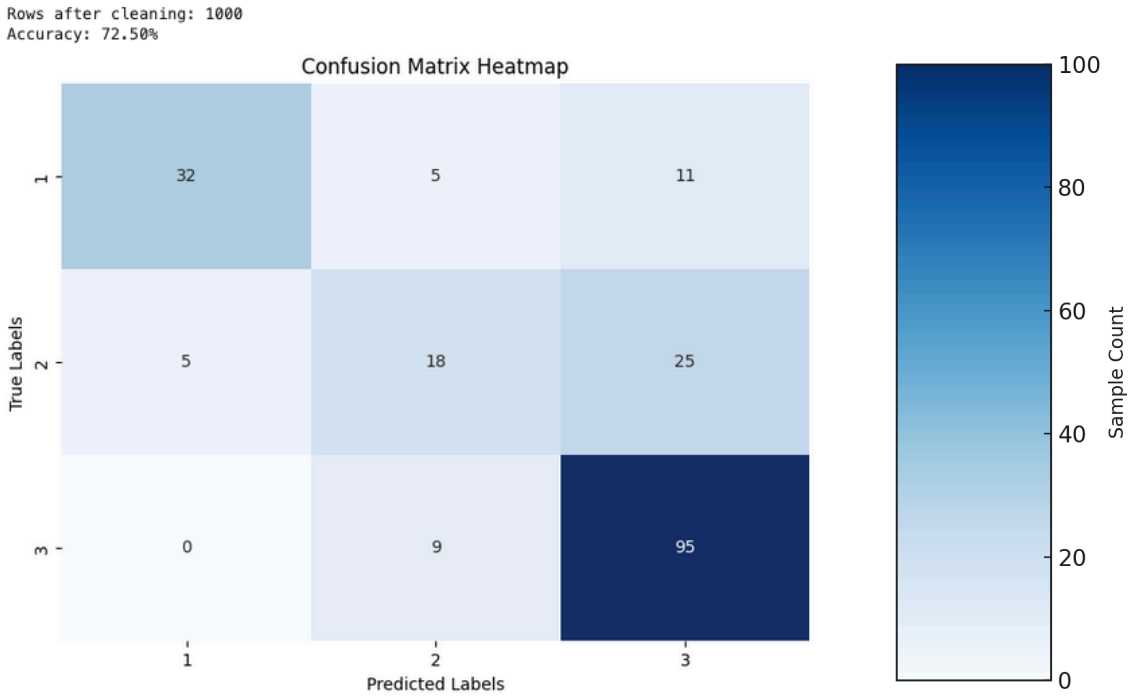

Fig. 2 Confusion Matrix Heatmap (After shortening to 3 bins)

To convert the problem to a classification task, I grouped the HbA1c values into three bins : Normal (HbA1c < 5.7), Prediabetic (5.7 ≤ HbA1c < 6.5), and Diabetic (HbA1c ≥ 6.5). They were chosen because they follow traditional medical diagnostic thresholds and give a relatively even split of the data over classes, preventing class imbalance during training. I used three classification models. The Decision Tree Classifier [6] because of its interpretability and ability to learn non-linear boundaries between decisions. The Random Forest Classifier [5] was used to prevent overfitting by averaging the predictions of numerous decision trees to generalize well. The K-Nearest Neighbors [KNN, 7] classifier was used because it predicts on the basis of feature similarity, which performs very well after normalization. The data was divided into training and test dataset with an 80/20 ratio. Accuracy scores and confusion matrices were used to evaluate the performance of models. Lastly Hyperparameter tuning was conducted to optimize models performance based on parameters like n_estimators, max_depth, and random_state for tree based classifiers and n_neighbors for KNN. I also considered XGBoost as a potential model given its strong track record on structured datasets, but due to storage and computational constraints it was not implemented in this study.

Finally, I implemented a Voting Classifier [8] that combined the prediction of the Random Forest, Decision Tree, and KNN models. The ensemble technique enhanced the performance and achieved a higher accuracy of 72.5%, which means that employing an ensemble of multiple models is likely to provide more accurate results than the implementation of a single model. All algorithms were implemented in Python using Scikit-learn [11].

RESULTS

Regression models demonstrated limited ability to make accurate HbA1c predictions. Linear Regression [2, 12] had a poor linear relationship of predictors and HbA1c, with only an R^2 of 0.16 and an RMSE of around 1.42. The Decision Tree Regressor [3] did slightly better, with an R^2 of 0.21 and an RMSE of 1.34, but Random Forest Regressor [4] was best in regression, with an R^2 of 0.29 and an RMSE of 1.27. These relatively low R^2 and moderate errors indicate that the dataset had too little depth and variety in features to enable making continuous predictions that were accurate, and might be due to having limited clinical variables or biased feature distributions.

Reframing the problem as a classification problem enhanced predictive accuracy to a large degree. Random Forest Classifier [5] had optimum accuracy of up to 68.1%, followed by Decision Tree Classifier at 55.1%, and K-Nearest Neighbors (KNN) Classifier at 52.7%. Perusal of the Random Forest confusion matrix indicated that the algorithm performed well in detecting diabetic cases, and errors were mainly at boundary locations near normal and prediabetic classes. This represents a suggestion that the algorithm had identified overall patterns well but not boundary cases near threshold regions.

Hyperparameter tuning increased the accuracy of the model as well. The best performance of the Random Forest Classifier [5] occurred at 100 estimators and 12 max depth; higher depths allowed for greater overfitting likelihood, lower depths decreased the model's predictability. Classification models were in general preferable for this set of data because HbA1c range identification as categories always remained a more reliable task than numerical value prediction of the provided features. Even though these findings demonstrate the ability of

machine learning classifiers to differentiate HbA1c classes, they fall short in some ways. The dataset's small feature set, and the absence of behavioral or longitudinal data definitely limited performance, especially in boundary scenarios. Limitations of this sort offer scope for future research to increase the set of variables, incorporate temporal trend features, and to test on heterogenic, larger populations.

CONCLUSION

In this study, I predicted HbA1c levels using machine learning models from a public dataset [1]. This analysis yielded moderately disappointing results when using regression models such as Linear Regression, Decision Tree Regressor, and Random Forest Regressor [4] because HbA1c as a continuous variable had a poor fit according to models calibrated on binary values so I reframed the analysis as a class based prediction problem for HbA1c levels presented in a binary format. Changing the continuous prediction problem of HbA1c into a class-based prediction provided better prediction across the three final candidate models over prior models and approaches. In case analysis, the Random Forest model yielded the best model accuracy at 68.1%, while the Voting Classifier yielded the highest outcomes at 72.5% [8]. Overall, results demonstrated that class-based prediction is supported in limited-feature datasets such as HbA1c levels, while ensemble classifiers yield further improvement in accuracy [13]. What is important beyond errors and seeming technical specific analyses, was the opportunity for machine learning approaches to impact early detection and management of diabetes as support in situations where active laboratory means and resources are limited. Expansion of dataset size, perspective on incorporating behavioral and longitudinal approaches and validation on a more diverse population, should be the focus of future research to improve clinical relevance and impacts.

ACKNOWLEDGEMENT

I would like to acknowledge the guidance and mentorship of Faisal Qureshi for his continuous guidance throughout the course of my research.

REFERENCES

[1] AravindPCoder. (2023, November 18). Diabetes dataset. Kaggle.

https://www.kaggle.com/datasets/aravindpcoder/diabetes-dataset?resource=download

[2] Scikit-Learn Developers. (2025). *Linear Regression documentation.* Scikit-Learn.

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[3] Scikit-Learn Developers. (2025). *Decision Tree Regressor documentation.* Scikit-Learn.

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

[4] Scikit-Learn Developers. (2025). *Random Forest Regressor documentation.* Scikit-Learn.

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[5] Scikit-Learn Developers. (2025). *Random Forest Classifier documentation.* Scikit-Learn.

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[6] Scikit-Learn Developers. (2025). *Decision Tree Classifier* documentation. Scikit-Learn.

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[7] Kartik. (2025, August 23). *K-nearest neighbors (KNN).* GeeksforGeeks.

https://www.geeksforgeeks.org/machine-learning/k-nearest-neighbours/

[8] Scikit-Learn Developers. (2025). *Voting Classifier documentation.* Scikit-Learn.

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html

[9] Alhassan, Zakhriya, et al. “Improving Current Glycated Hemoglobin Prediction in Adults: Use of Machine Learning Algorithms with Electronic Health Records.” *JMIR Medical Informatics*, U.S. National Library of Medicine, 24 May 2021,

pmc.ncbi.nlm.nih.gov/articles/PMC8185616/.

[10] Tao, X., Jiang, M., Liu, Y., Hu, Q., Zhu, B., Hu, J., et al. (2023, September 30). *Predicting three-month fasting blood glucose and glycated hemoglobin changes in patients with Type 2 diabetes mellitus based on multiple machine learning algorithms.* Scientific Reports.

https://doi.org/10.1038/s41598-023-43240-5

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 12, 2825–2830.

https://doi.org/10.48550/arXiv.1201.0490

[12] GraphPad by Dotmatics. (n.d.). *Linear regression calculator.*

https://www.graphpad.com/quickcalcs/linear1/

[13] Tablas-Mejia, I. (2025). *Conclusion section for research papers*. San José State University Writing Center.

https://www.sjsu.edu/writingcenter/docs/handouts/Conclusion%20Section%20for%20Research%20Papers.pdf