Advancing Diabetes Prediction: A Nuanced Six-Class Classification System and Risk Factor Interactions Investigation

Shengyuan Zhang

Statistics-data science track, Cornell University, Ithaca, NY 14850, US
sz663@cornell.edu

Abstract. This study advances diabetes prediction by introducing a nuanced, six-class classification system and examining the interaction effects of various risk factors. Rather than the traditional binary classification, this research proposes six distinct diabetes classes: normal, pre-diabetic, diabetic under control, diabetic fair control, diabetic poor control, and diabetic very poor control. These classes, derived from Hemoglobin A1c (HbA1c) and blood sugar levels, provide healthcare professionals and patients with a more comprehensive understanding of the disease. Machine learning algorithms, including Logistic Regression, Random Forest, and Dense Neural Network (DNN) for binary classification, and Random Forest, Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), CatBoost, and DNN for six-class classification, were employed to compare accuracy rates. Risk factors such as Body Mass Index (BMI), age, blood sugar level, and HbA1c level were categorized, and their interaction effects were evaluated using conditional entropy and visualized with hierarchical clustering, dendrograms, and heatmaps. The findings reveal that multi-class diabetes prediction can achieve comparable accuracy to binary classification when HbA1c and fasting blood sugar levels are accurately measured. Moreover, the investigation into interaction effects yields valuable insights into the heightened risk associated with the combination of major risk factors.

Keywords: Machine Learning, Diabetes Prediction, Interaction Effects Analysis

1 Introduction

Diabetes, particularly Type 2 Diabetes Mellitus (T2DM), is a significant global health concern with a rising prevalence, especially in middle- and low-income countries. The disease is a major cause of blindness, kidney failure, heart attacks, stroke, and lower limb amputation. Various risk factors such as genetic disposition and body mass index. contribute to the development of diabetes. In this context, machine learning has emerged as crucial tools in the early detection and management of diabetes. These technologies can analyze large amounts of data and identify patterns, unveiling patterns that may elude human perception. For instance, machine learning can analyze

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_71

patient data to predict who is at risk of developing diabetes, allowing for early intervention and potentially preventing the onset of the disease [1]. They can be used to improve the diagnosis of diabetes and its complications, such as diabetic retinopathy [2]. Early diagnosis is crucial for patients to control diabetes and prevent complications. As a result, diabetes has garnered considerable attention within the field of artificial intelligence-assisted medical diagnosis.

Machine learning algorithms have been extensively used in predicting diabetes, with each study employing various techniques to enhance the predictive power of their models. One study utilized a variety of machine learning models, including models e.g. logistic regression and decision tree trained on a large dataset of patient records, showing promising results in predicting the co-occurrence of diabetes and cardiovascular diseases [3]. Another research used an ensemble of machine learning models, including e.g. Random Forest and AdaBoost trained on the Pima Indian Diabetes Dataset [4]. In a distinct investigation, a variety of machine learning techniques were employed, including decision trees, random forests, and gradient boosting machines [5]. A recent study employed a deep learning model, specifically a convolutional neural network (CNN), to predict diabetes using retinal fundus images [6]. Furthermore, a study used a fused machine learning approach for diabetes prediction, which includes Support Vector Machine (SVM) and Artificial Neural Network (ANN) models [7].

This study, however, aims to delve deeper into the interaction effects among these predictors. By employing hierarchical clustering and dendrogram heat maps, the conditional probability of diabetes given other variables can be visualized. Furthermore, by fusing two variables, such as age and BMI, the conditional probability of diabetes given these combined features can be explored. This approach can uncover more information about interaction effects, potentially enhancing the predictive power of the models.

In addition to exploring interaction effects, this research also aims to provide a more nuanced classification of diabetes. Instead of the classical binary classification (diabetes or not), this study proposed a system that classifies diabetes into six specific categories: normal, pre-diabetes, diabetes under control, diabetes fair control, diabetes poor control, and diabetes very poor control. These labels, derived from a combination of HbA1c and fasting blood sugar levels, provide patients and healthcare professionals with more detailed information about the disease, facilitating better decision-making and more personalized treatment strategies.

The importance of this research lies in its potential to revolutionize the understanding and prediction of diabetes. By exploring interaction effects and providing a more nuanced classification of diabetes, a more comprehensive understanding of the disease and its risk factors can be obtained. This, in turn, can lead to more effective prevention and treatment strategies, ultimately improving the quality of life for individuals affected by diabetes.

2 Method

2.1 Data Description

The first dataset termed as Dataset A used in this study [8] is sourced from Kaggle and primarily derived from Electronic Health Records (EHRs). EHRs are digital versions of patient health records that contain information about their medical history, diagnosis, treatment, and outcomes. The data in EHRs is collected and stored by healthcare providers, such as hospitals and clinics, as part of their routine clinical practice. However, it is crucial to acknowledge the presence of limitations within this dataset. Despite its expansive sample size, the data within the EHRs exhibits imperfections in terms of cleanliness. Some subjects demonstrate a disparity between exceptionally elevated blood glucose levels and relatively diminished hemoglobin levels. Moreover, numerous subjects were falsely classified into non-diabetic despite they are actually diabetic.

The other dataset termed as Dataset B used [9], also sourced from Kaggle, consists of over 100 patient records collected from a variety of sources, including medical records, surveys, and interviews. This dataset is preferred due to its cleanliness. Data were collected with fasting blood glucose to prevent errors, and the diagnosis column only indicates whether the subject was formally diagnosed or not. It does not reflect whether the subject has diabetes. However, a limitation of this dataset arises from its relatively small sample size, comprising less than 150 observations.

Both datasets provide valuable insights into the factors associated with diabetes and can be used to develop predictive models. However, the limitations of each dataset, such as the cleanliness of the data in Dataset A and the small sample size in Dataset B, should be taken into consideration when using these datasets for investigation. Table 1 provides the overview of the parameters in two employed datasets.

Table 1. The parameters in both datasets.

Parameters	Dataset A	Dataset B
Age	Yes	Yes
Gender	Yes	Yes
BMI (Body Mass Index)	Yes	Yes
Hypertension	Yes	No
Heart Disease	Yes	No
Smoking History	Yes	Yes
HbA1c Level	Yes	Yes
Blood Glucose Level	Yes	Yes
Diabetes Status	Yes	No
Diagnosed by medical professional(s)	No	Yes
Blood Pressure	No	Yes
Fasting Blood Sugar (FBS)	No	Yes
Family History of Diabetes	No	Yes
Diet	No	Yes
Exercise	No	Yes

2.2 Data Preprocessing

In this study, a comprehensive data preprocessing pipeline was meticulously implemented on a diabetes dataset to ensure its suitability for machine learning analysis. The initial phase involved discretizing continuous variables such as Age, Body Mass Index (BMI), Hemoglobin A1c (HbA1c), and Fasting Blood Sugar (FBS) into categorical bins. This transformation is crucial in facilitating the application of certain machine learning algorithms.

A significant part of the preprocessing involved the creation of a new column, 'diagnosis_combined', by merging the 'Diagnosis', 'FBS', and 'HbA1c' columns. This was not merely a reassignment of labels but a strategic move to transform the problem from a binary classification to a multiclass classification. The new labels, including 'Normal', 'Prediabetic', 'Diabetic-Fair Control', 'Diabetic-Poor Control', and 'Diabetic-Very Poor Control', were assigned based on the combined information from the three original columns. This transformation allowed for a more nuanced understanding of the diabetes condition, moving beyond a simple binary diagnosis and providing a more detailed classification that reflects the varying degrees of severity in diabetes. Following the reassignment process, it is crucial to take out blood glucose level and HbA1c level columns to ensure the predictive model does not have access to information that would not be available in a real-world prediction scenario without medical examinations. This data leakage directly causes model overfitting and lack of generalizability and interpretability of other independent variables.

To address class imbalance in the target variable, an oversampling technique was utilized using the RandomOverSampler from the imbalanced-learn library [10], mitigating model bias towards the majority class. Lastly, categorical variables were encoded using two techniques, namely LabelEncoder and One-Hot Encoding. While LabelEncoder converts each category into a unique integer, One-Hot Encoding transforms each category value into a new column and assigns a binary value of 1 or 0.

2.3 Interaction

For the interaction effects, a combination of hierarchical clustering and conditional entropy was employed to explore the interaction among parameters in predicting diabetes with severity. Hierarchical clustering, a method of cluster analysis, was used to build a hierarchy of clusters, visually represented by a dendrogram. This tree-like diagram, when combined with a heatmap, heatmap's color gradient, ranging from red to blue, was determined by the conditional probabilities, effectively illustrating the magnitude of the relationship between variables. This approach revealed hidden patterns and dependencies that might not be readily apparent in a simple univariate visualization.

Complementing this, conditional entropy was used to quantify the amount of information needed to describe the outcome of a random variable given the value of another variable. This measure was used to rank the parameters based on their explanatory power over the variance in diabetes severity classes. A lower conditional entropy indicates a stronger interaction effect, as knowing the state of one variable reduces the uncertainty of the other. With fused variable such as ‘age_BMI’ which intuitively combines age and BMI parameters, to understand their joint influence on diabetes severity. This approach recognizes that the impact of a single risk factor on diabetes may change depending on the level of another risk factor.

The integration of hierarchical clustering and conditional entropy in this study underscores the importance of exploring interaction effects in understanding diabetes risk factors. By revealing the complex interplay among risk factors, this approach provides a more nuanced understanding of diabetes risk, which can inform more effective prevention and treatment strategies. This research serves as a valuable guide for healthcare professionals and patients, highlighting the most significant factors to monitor in managing diabetes risk.

2.4 Prediction Algorithms

In this study, several machine learning algorithms was employed, including CatBoost, an algorithm specifically designed to handle categorical variables, and Deep Neural Networks (DNNs), which are capable of modeling complex non-linear relationships. Various tree ensemble methods such as Random Forest, LightGBM, and XGBoost are also used. Random Forest operates by constructing multiple decision trees and outputting the class that is the mode of the classes or mean prediction of the individual trees. LightGBM and XGBoost are gradient boosting frameworks that use tree-based learning algorithms, designed to be efficient and capable of handling large-scale data. These algorithms were chosen for their proven effectiveness and versatility in handling various types of data and predicting complex outcomes.

3 Results and Discussion

3.1 Algorithm Prediction

The prediction results varied significantly between the two datasets shown in Table 2. For Dataset A, accuracy rates ranged from 38.98% to 51.22%. In contrast, the same machine learning algorithms applied to Dataset B yielded accuracy rates between 96.5% and 100%. Notably, DNNs consistently performed better on Dataset A. However, for Dataset B, LightGBM, XGBoost, and DNN all achieved 100% accuracy. This suggests that with sufficiently clean data, DNNs could potentially outperform algorithms such as XGBoost and LightGBM.

Interestingly, the accuracy rates for Dataset B did not decrease when predicting five multi-class categories instead of binary classes. In contrast, the highest accuracy rate for Dataset A dropped from 96% to 51.22% when predicting multi-class categories. This indicates that with clean and sufficient data, it is possible to provide a nuanced classification of diabetes severity without sacrificing prediction accuracy.

Table 2. The prediction performance of various algorithms on Dataset A and Dataset B.

Algorithms	Dataset A Accuracy rate	Dataset B Accuracy rate
CatBoost	38.98%	96.5%
Random	48.98%	97.5%

Forest		
LightGBM	44.93%	100%
XGBoost	45.70%	100%
DNN	51.22% (early stopping + regularization)	100% (with and without early stopping)

3.2 Interaction Effects

The significance of interaction effects is underscored by the compelling visualizations generated in this study and corresponding results are presented in Table 3, Table 4, Fig. 1, Fig. 2, Fig. 3 and Fig. 4. Both datasets incorporate key variables such as age, gender, body mass index (BMI), and smoking history. Age and BMI are identified as some of the most informative variables based on the conditional entropy ranking, a result that is far from arbitrary. This entropy ranking is further illustrated in the accompanying table.

The uniformity of hierarchical clustering visualizations, including dendrograms and heat maps, across both datasets is remarkable. Several variables are intuitive and self-explanatory, underscoring their significance in predicting diabetes. Despite the challenges associated with data cleanliness in HbA1c and blood glucose levels, this study suggests that variables such as age, BMI, gender, smoking history, and blood pressure (along with related diseases) are the most informative and crucial in predicting diabetes.

The interaction effect between two risk factors becomes evident when they are analyzed in conjunction. From the single-variable hierarchical clustering analysis, it is observed that subjects with high BMI, particularly those in the severely obese category, are highly likely to have diabetes. However, when the diagnosis is combined with BMI from Dataset B, an intriguing pattern emerges among subjects within the same body weight category. Specifically, those not formally diagnosed are significantly more likely to fall into the 'very poor control' diabetic group. In contrast, individuals of the same weight group who were formally diagnosed are more likely to be categorized into the 'poor control' and 'fair control' diabetic groups. Similarly, Age has been consistently identified as a strong predictor for diabetes. However, within the same age group, different BMI categories markedly influence the conditional probability of varying diabetes severity. Each combination of risk factors offers unique insights, shedding light on the cumulative impact these factors can exert on the severity of diabetes. This nuanced understanding can guide targeted interventions and inform predictive models.

Table 3. Conditional Entropy of Diabetes Given One Other Variable

	Dataset A	Dataset B
CE(diabetes | gender)	2.1777	1.9164
CE(diabetes | age)	2.0241	1.4775
CE(diabetes | hypertension)	2.1827	n/a
CE(diabetes | heart disease)	2.1856	n/a
CE(diabetes | smoking history)	2.1226	2.3019

CE(diabetes | BMI)	2.1028	1.1345
CE(diabetes | blood pressure)	n/a	1.2949
CE(diabetes | family history of diabetes)	n/a	2.3118
CE(diabetes | diet)	n/a	2.3019
CE(diabetes | exercise)	n/a	2.3019
CE(diabetes | diagnosis)	n/a	2.2255

Table 4. Conditional Entropy of Diabetes Given Fused Two Variables (lowest ten conditional entropy only)

Condition	Dataset A	Dataset B
CE(diabetes | gender + smoking history)	2.1117	1.8984*
CE(diabetes | heart disease + BMI)	2.1004	n/a
CE(diabetes | hypertension + BMI)	2.0990	n/a
CE(diabetes | gender + BMI)	2.0940	0.7632
CE(diabetes | smoking history + BMI)	2.0565	1.0989
CE(diabetes | age + heart disease)	2.0222	n/a
CE(diabetes | age + hypertension)	2.0212	n/a
CE(diabetes | gender + age)	2.0148	1.0641
CE(diabetes | age + smoking history)	1.9892	1.4477*
CE(diabetes | age + BMI)	1.9786	0.6174
CE(diabetes | BMI + Blood Pressure)	n/a	0.7778
CE(diabetes | Age + Blood Pressure)	n/a	0.7632
CE(diabetes | Gender + Blood Pressure)	n/a	1.0276
CE(diabetes | BMI + Diagnosis)	n/a	1.0374
CE(diabetes | BMI + Diet)	n/a	1.0989
CE(diabetes | BMI + Exercise)	n/a	1.0989

*not lowest 10 ranking entropy but available

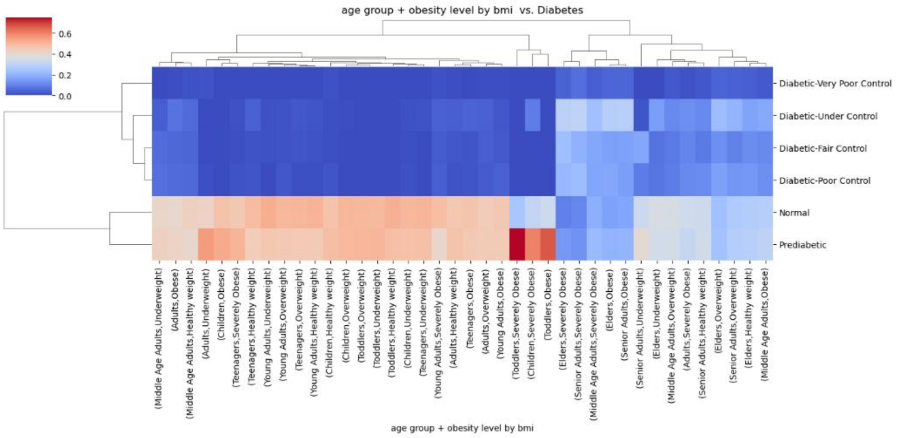

Fig. 1. Hierarchical Clustering with Conditional Probability of Diabetes vs. Fused Variable Age and BMI from Dataset A (Photo/Picture credit: Original).

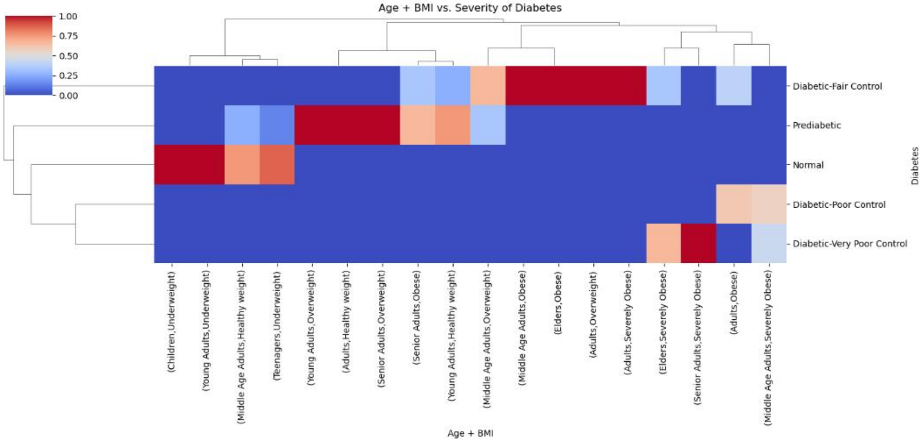

Fig. 2. Hierarchical Clustering with Conditional Probability of Diabetes vs. Fused Variable Age and BMI from Dataset B (Photo/Picture credit: Original)

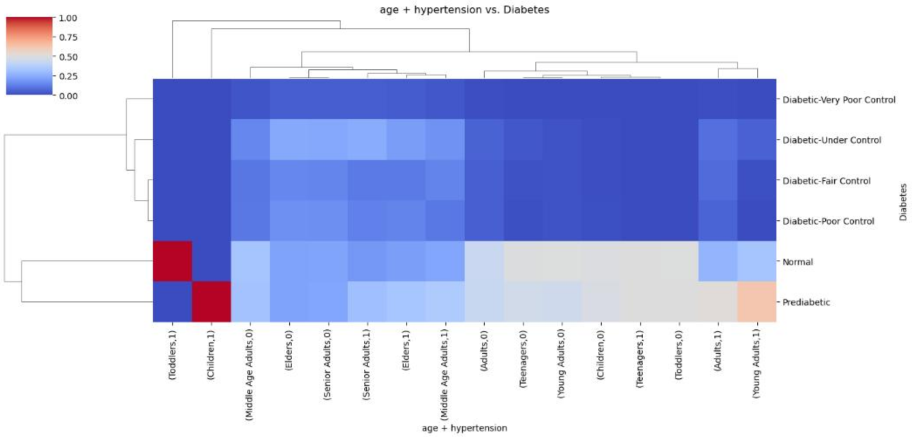

Fig. 3. Hierarchical Clustering with Conditional Probability of Diabetes vs. Fused Variable Age and Hyperextension from Dataset A (Photo/Picture credit: Original)

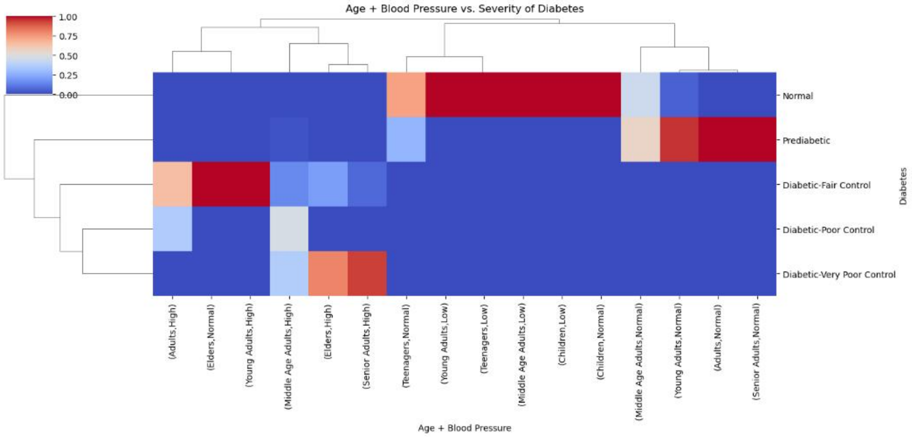

Fig. 4. Hierarchical Clustering with Conditional Probability of Diabetes vs. Fused Variable Age and Blood Pressure from Dataset B (Photo/Picture credit: Original)

4 Conclusion

This study advances the understanding of diabetes by introducing a multi-class classification and examining the interaction effects of various risk factors. Using conditional entropy on two distinct datasets, risk factors were ranked based on their explanatory power, and pairs of risk factors were combined to reveal insightful interaction effects. Hierarchical clustering, dendrograms, and heatmap visualizations were employed to illustrate these effects, providing healthcare professionals with a deeper understanding of how combined risk factors can exacerbate diabetes in certain patients. Various machine learning algorithms, including CatBoost, Random Forest, XGBoost, LightGBM, and DNN, were utilized to enhance the performance of the multi-class classification. The results indicate that with accurate fasting blood sugar levels, as in Dataset B, multi-class classification can achieve similar accuracy rates to binary classification. However, despite the large sample size of Dataset A, the algorithms did not perform well. With clean and accurate fasting blood sugar Dataset B performed exceptionally well, despite containing fewer than 150 samples. Future research should focus on larger datasets with accurate fasting blood sugar and HbA1c levels to further study the accuracy of multi-class diabetes classifications.

References

1. Odedra, D., Subir, S., and Ambarish S. V., Computational intelligence in early diabetes diagnosis: a review. The review of diabetic studies: RDS 7.4 (2010): 252.
2. Sikder, N., et al.: Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. Symmetry 13.4 (2021): 670.
3. Abdalrada, A. S., et al.: Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases: a retrospective cohort study. Journal of Diabetes & Metabolic Disorders 21.1 (2022): 251-261.
4. Dutta, A., et al.: Early prediction of diabetes using an ensemble of machine learning models. International Journal of Environmental Research and Public Health 19.19 (2022): 12378.
5. Larabi-Marie-Sainte, S., et al.: Current techniques for diabetes prediction: review and case study. Applied Sciences 9.21 (2019): 4604.
6. Shin, J. Y., et al.: Development of various diabetes prediction models using machine learning techniques. Diabetes & Metabolism Journal 46.4 (2022): 650-657.
7. Ahmed, U., Issa, G. F., Khan, M. A., Aftab, S., Khan, M. F., Said, R. A., ... & Ahmad, M.: Prediction of diabetes empowered with fused machine learning. IEEE Access, 10, 8529-8538 (2022).
8. Mustafa, M.: Diabetes prediction dataset, Apr. (2023). https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset
9. Mandala, S.: Easiest Diabetes Classification Dataset, May (2023). https://www.kaggle.com/datasets/sujithmandala/easiest-diabetes-classification-dataset
10. Hai, W. A. N. G., et al.: Comparative Study of Oversampling and Ensemble Learning Methods in Software Defect Prediction. Computer and Modernization 06 (2020): 83.