Machine Learning-based Classification Model for Diabates*

Yuhe Fu[1], Xiaoyuan Zhang[1], and Huiyuan Zhou[1]

Duke Kunshan University, Kunshan 215316, China

Abstract. The increasing prevalence and severity of diabetes have caused both financial and physical burdens to millions of the population around the globe. As one of the most common chronic diseases, diabetes has become the top public health priority in many countries. Despite numerous scientific studies on diabetes, the exact pathology of it is yet to be discerned. Thus, an effective screening and diagnostic tool is needed for early prevention and effective treatment. In this work, we aimed to build a classification model to predict if a subject has no ignorable risk of diabetes, risk of pre-diabetes, or risk of diabetes based on the largest continuous nationwide telephone survey: The Behavioral Risk Factor Surveillance System (BRFSS) 2015 data. After exploratory data analysis, most of the respondents have no diabetes (83%), while only few have pre-diabetes (2%) or diabetes (15%). When one class is greatly overrepresented in another, it is known as class imbalance, and this can frequently result in skewed model performance and lower prediction accuracy. Given this challenge, we also attempted to tackle the underlying class imbalance issue in our dataset by applying oversampling, undersampling, and the Synthetic Minority oversampling Technique (SMOTE) methods. We experimented with a number of popular machine learning methods for classification including logistic regression, support vector machine, and AdaBoost. With no manipulation of the original dataset, the logistic regression with none penalty and balanced class weights achieved the best macro-level recall (51.39%) on the test set, which presents a promising machine learning-based tool for mass screening.

Keywords: Diabetes · Machine learning · Class imbalance

1 Introduction

Being one of the most prevalent diseases worldwide, diabetes is currently affecting millions of people annually and it has caused considerable impact on the national economy, rendering it the top public health priority in many countries. In 2011, it was estimated that 366 million people had predominantly type-2 diabetes, and this figure is anticipated to grow to 552 million by 2030 [1]. Diabetes is characterized by patients' inability to regulate the level of glucose after food intake, possibly due to β cell dysfunction and insulin resistance which negatively

* Supported by Duke Kunshan University, Kunshan 215316, China

influence insulin's secretion and its sensitivity respectively [2]. Previous evidence has shown that diabetes may lead to a variety of adverse health outcomes including impaired vision [3], dementia [4], reduced life expectancy [5], declined quality of life [6], etc. So far, there has not been a cure for diabetes, however, it has been revealed that healthy lifestyles such as physical activity, healthy eating, and early preventive strategies can reduce the risk of diabetes [7]. Early detection of pre-diabetes or diabetes is crucial to those at higher risk as the predictive results can encourage changes and improvement in lifestyles, leading to reduced or eliminated risk of further physical and financial burdens [8]. Currently, early diagnosis and screening tools include the collection of biomarkers such as Cystatin C, Homocysteine, NGAL, NGAL, etc [9]. However, the reliability of diagnosis based on only a subset of the biomarkers remains a question, not to mention the great cost and time induced by the collecting and analyzing process [10].

To improve the accuracy of diabetes screening and lower the cost of such screening campaigns, machine learning-based techniques are gaining attention for their successful applications in cancer prognosis, medical image classification, and drug development [11–13]. Machine learning models for medical diagnosis, which in our case is the classification of the diabetes type, come with several compelling advantages. Firstly, machine learning models, once trained, could be applied to analyzing vast amounts of data in a short time, which is extremely favorable in terms of mass screening. Additionally, a well-designed machine learning model will be able to generate the predicted result rapidly, reducing the time cost of early diagnosis initiatives in a clinical setting. Lastly, machine learning models are often capable of depicting complex and non-linear relationships between predictors and the outcome in real-life applications and highlighting risk factors that may not receive adequate attention [14]. Despite the robustness of machine learning algorithms in clinical practice, many machine learning-based attempts have failed due to reasons ranging from low-quality data to low generalizability from the training set to the test set [15]. As its name suggests, machine learning algorithms heavily rely on the learning process to make accurate and reliable predictions. In the realm of real-world data, particularly in classification tasks, class imbalance emerges as a prevalent issue. Chronic diseases like diabetes or rate diseases like thyroid disorders often account for a significantly low proportion among collected datasets, leading to skewed data distribution and underfitting for the minority class. Even popular machine learning models including random forest (RF) which typically exhibit extraordinary performance in various scenarios often fail in the minority class classification when trained on imbalanced data as most of them assume an equal misclassification cost and balanced class distribution [16, 17].

In this work, our goal is to build a reliable machine learning model to classify respondents as having no diabetes, pre-diabetes, and diabetes while experimenting with techniques to tackle class imbalance: oversampling, undersampling, and the Synthetic Minority Oversampling Technique (SMOTE). Our ultimate objective is to suggest a combination of a machine learning model and a class-imbalance solution that could yield optimal performance.

2 Background

To understand the state-of-the-art progress in machine learning-based classification models for diabetes, we searched on PubMed using the keywords "BRFSS" + "diabetes" + "machine learning". The article that matches well with our work is one titled: "Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques" by Xie et al. [18], which is based on BRFSS 2014 data. In this article, the authors excluded the pre-diabetes category from the original dataset, and they applied the SMOTE technique after noticing the class imbalance in the training set. Using the accuracy, sensitivity (recall) and specificity (true negative rate), the authors attempted neural network, logistic regression, SVM with linear, rbf, and polynomial kernel, random forest, naive bayes, and decision tree. The highest sensitivity was given by the decision tree model (51.6%). Although the neural network has the best accuracy (82.4%) and AUC (79.5%), it performed the worst in terms of recall. SVM with different kernels did not perform as well as expected either and was ranked even after multivariate logistic regression.

In addition to searching for articles that used BRFSS data, we also looked for articles reporting machine learning models constructed with class imbalance solutions. In Dagliati et al.'s (2017) article [19], the authors built several machine learning models and compared their performance with oversampled and original datasets. All models' AUC improved drastically after the oversampling strategy was applied. It is noteworthy that logistic regression's AUC was leading when no oversampling was applied, and its AUC was still among the top 3 after oversampling was applied, following random forest and SVM.

Specifically regarding techniques for class imbalance, Sadeghi et al.'s (2022) article [20] introduced 7 potential solutions for class imbalance: 1) **threshold moving** that moves the usual 0.5 threshold to an appropriate value 2) **cost-sensitive learning** that assigned higher weights to minority class for the loss function 3) **repeated edited nearest neighbors** which is a type of undersampling method which is robust to noise and redundant data 4) **one sided selection** which is an undersampling method that constructs a subset of the data consisting of mainly the minority class 5) **synthetic minority oversampling technique** (SMOTE) which generates synthetic samples by n-nearest neighbors 6) **SVM-SMOTE** which only oversamples instances from minority class that are at the borderline. Borderline instances are determined by the support vectors of the SVM algorithm 7) **hybrid approach** that comprises a combination of under- and oversampling techniques. ENN-SMOTE works by undersampling the majority class by the edited nearest neighbors and oversampling instances from the minority class by SMOTE.

3 Problem

3.1 Domain

As one of the most prevalent chronic diseases, diabetes has caused millions of people around the globe financial and physical burdens. Building an accurate classification model to classify participants with no diabetes, prediabetes and diabetes contributes to prompt diagnosis, early preventive strategies, and effective treatment plans. Various health-related risk factors such as blood pressure, cholesterol, and body mass index (BMI) might influence the chance of an individual developing diabetes or pre-diabetes.

3.2 Problem Statement

The problem that we are going to address in this paper is the prediction of 3 different categories of diabetes based on a selected set of features. The objective is not only concerning overall accuracy but also the recall of the minority class where samples in the data set are significantly fewer than the majority class of no diabetes.

3.3 Formal Definition

Let $X = \{x_1, x_2, ..., x_{253,680}\}$ be the data set consisting 253,680 samples, where each sample $x_i = (x_{i1}, x_{i2}, ..., x_{i21})$ is a vector of features representing 45 different attributes of one sample. Features in the vector include attributes such as blood pressure, cholesterol, and body mass index (BMI). The ultimate goal is to learn a function $f : R^{45} \rightarrow R$ from the training data, where the prediction error on the testing data will be minimized. The output of the function f will be a discrete label value $\hat{y}_i$, which is the predicted label of the sample with features x_i. The label value of 0 is for no diabetes, 1 is for pre-diabetes, and 2 is for diabetes.

3.4 Objective

In general, the objective is to minimize the testing error. One of the most common choices for measuring the testing error is the accuracy score:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where the notation TP means true positive, TN means true negative, FP means false positive and FN means false negative. However, the dataset that our project utilized has a significant class imbalance issue, where the majority class (class 0) has 190,055 samples, which accounts for nearly 83% of all samples. For a dataset with an imbalanced target variable distribution, a high accuracy score might be misleadingly generated by a great performance on the classification of the majority class (class 0) while having poor performance on the minority class (class 1 & class 2). Therefore, the accuracy score may not be a suitable

criterion to evaluate the performance of the trained machine learning models for our classification task. What's more, for the classification of different statuses of diabetes, the cost of a false positive (classify a sample without diabetes as having prediabetes or diabetes) and false negative (classify a sample with diabetes or prediabetes as having no diabetes) is not equivalent, where the former one can be tolerated and the latter one may lead to delayed hospitalization or medication.

To best identify individuals at higher risk without missing potential patients, the macro recall score is considered in the present study:

$$Recall = \frac{TP}{TP + FN}$$

Considering we are tasked with a multi-class classification problem, the recall score can be extended to the macro recall:

$$Macro - Recall = \frac{\sum_{i=1}^{n} Recall_i}{n}$$

As the formula above suggests, the macro recall score is calculated by taking the unweighted average of the recall score of each class. Compared to the accuracy score, the performance of the model in each class has an equal contribution to the evaluation of the overall performance of the model on the testing data, which will alert the poor performance of a certain class.

4 Method

4.1 Data source

The data used in this work was retrieved from Kaggle based on the Behavioral Risk Factor Surveillance System (BRFSS) conducted by the US Centers for Disease Control and Prevention [25]. BRFSS was initiated in 1984 and it is now the largest continuous surveillance system in the world. The dataset contains 21 features, ranging from general health status (including hypertension, high cholesterol, BMI, etc) to income level. The target variable is categorized into 3 classes: class "0" represents the subjects with no diabetes, class "1" represents the subjects with pre-diabetes, and class "2" represents the subjects with diabetes. All responses were self-reported and recorded over the telephone.

4.2 Data pre-processing

Firstly, we checked the duplication in the dataset and deleted rows that were repeated. 23899 records out of 253680 (9.4%) were removed. Regarding the discrete valuables that have more than 2 categories in the dataset, we converted them into corresponding number of dummy variables which take the value of either 0 or 1. To reduce colinearity, the first category of each dummy variable was removed. For instance, the variable "Education" has 6 categories, and only 5

dummy variables: Education_2, Education_3.. Education_6 will be created. Furthermore, we noticed that features are not all on the same scale, which might lead to poor classification performance. Subsequently, standardization was performed for all features by subtracting the mean and dividing by the standard deviation:

$$Z = \frac{X - \mu}{\sigma}$$

where μ is the mean of that feature and σ is the standard deviation.

4.3 Sampling method

Three sampling methods were adopted to overcome the class imbalance issue. **Random undersampling**: instances from the majority class (i.e., no diabetes) were randomly sampled without replacement to match the number of the least represented class (i.e., pre-diabetes). **Random oversampling**: instances from the two minority classes (i.e., pre-diabetes and diabetes) were randomly sampled with replacement to match the number of the majority class. **SMOTE**: instances from the two minority classes were oversampled to match the number of the majority class according to the SMOTE algorithm depicted below:

Algorithm 1 Synthetic Minority Over-sampling Technique (SMOTE)

Input: Minority class samples S, Oversampling amount N, Number of nearest neighbors k
Output: Synthetic samples S_{synth}
Initialize S_{synth} as an empty set
for each sample $s \in S$ **do**
  Find k nearest neighbors of s in S, denote as N_s
  for $i = 1$ to N **do**
    Randomly select a neighbor $n \in N_s$
    Generate synthetic sample s_{new} by interpolating between s and n:
    $s_{\text{new}} = s + \lambda \times (n - s)$, where λ is a random number between 0 and 1
    Add s_{new} to S_{synth}
  end for
end for
return S_{synth}

The histogram of each predictor was also examined to ensure that the data distribution did not vary significantly after the sampling strategy was applied. As a result, no significant data distribution shift was observed (Appendix 1.).

4.4 Model development and evaluation

After data pre-processing, 60% of the data were randomly selected as the training set, and 40% were chosen as the test set in a stratified manner according to the

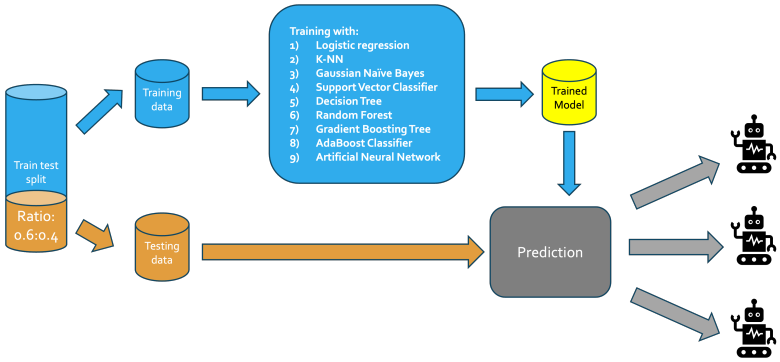

Fig. 1: Methodology of the Present Study

distribution in the target variable. Besides the Dummy Classifier where all samples are categorized as the majority class, Logistic Regression, K-Nearest Neighbour, Decision Tree, Gaussian Naive Bayes, Random Forest, Gradient Boosting Decision Trees, AdaBoost Classifier, Neural Network, and RBF SVM were applied after the train-test split. For each classification model above, it was trained and evaluated with different sampling strategies. The macro recall score for each combination of the classification method and sampling strategy was calculated, and the top 3 models were selected for hyperparameter tuning. Finally, the best model after hyperparameter tuning was selected as the best model, and the corresponding macro recall score was reported (Fig. 1.).

Table 1: Initial parameter setting of each model.

Model	**Initial Parameter**
Logistic Regression	penalty='l2', solver='newton-cg', class weight='balanced', max iter=1000
Gaussian Naive Bayes	default
Decision Tree	class weight='balanced'
Random Forest	n estimators=300, class weight='balanced'
Gradient Boosting	n estimators=300
AdaBoost	n estimators=300
K-Nearest Neighbors	n neighbors=5
Support Vector Machine	kernel='rbf', class weight='balanced'

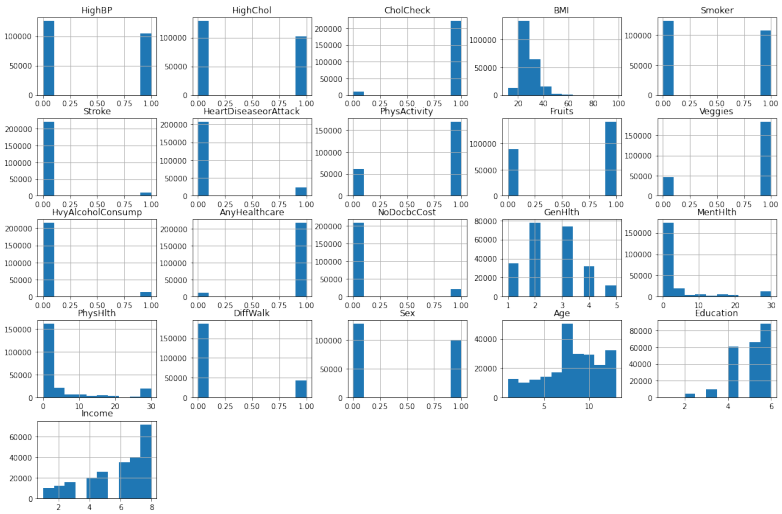

Fig. 2: Feature Distribution in the Original Data

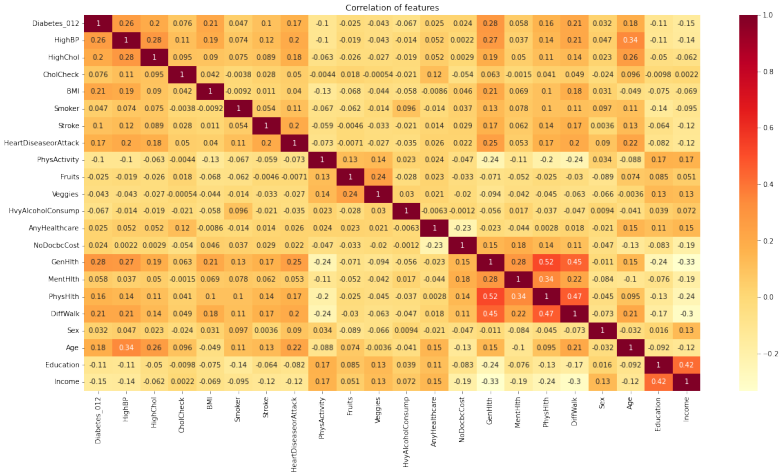

Fig. 3: Feature Correlation in the Original Data

5 Results

5.1 Exploratory Data Analysis

Prior to training our machine learning model, we conducted a comprehensive exploratory data analysis to gain insights into the dataset. Upon examining the histogram of each feature (Fig. 2.), it can be seen that most of the respondents are aged between 50 and 75 years old, who are more likely to suffer from diabetes. There are slightly more females than males in the BRFSS 2015, however, their difference is not significantly worrying. Notably, the continuous variable BMI exhibited a distribution pattern closely aligned with the normal distribution, indicating a good fit to the expected statistical properties. To prevent introducing collinearity in the features, we also checked the correlation between each pair of features (Fig. 3.). There is no obvious collinearity in the dataset, and the highest correlation coefficient is 0.52 between general health and physical health. Among all features, the high blood pressure correlates with the target variable the most, which might indicate that it is a discriminant predictor.

In addition, to visualize all the instances in 2D space, we conducted dimensionality reduction with principal component analysis (PCA). Through pairwise comparison (Fig. 4.), it can be observed that the normal group exhibits a more sparse distribution than the pre-diabetes and diabetes population while pre-diabetes and diabetes share similar distributions. This initial observation raises the possibility of potential challenges in accurately distinguishing between the pre-diabetes and diabetes groups in our subsequent analyses.

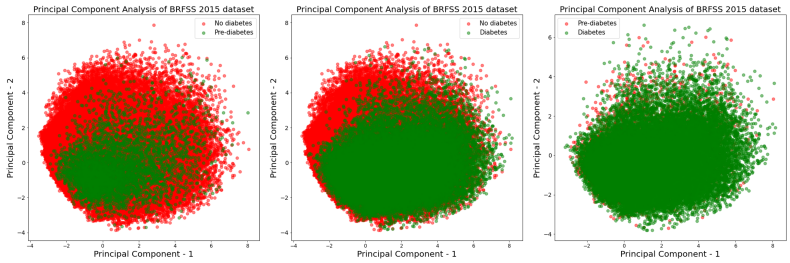

Fig. 4: Feature Correlation in the Original Data

5.2 Computational Results

The comprehensive comparison of various methods and models indicates the top three models based on macro recall values in different methods: for the original dataset, the top three models are Logistic Regression, Support Vector Machine (SVM), and Gaussian Naive Bayes (GNB); for the undersampled data, the top three models are AdaBoost, Gradient Boosting, and Logistic Regression; for the

oversampled data, the top three models are AdaBoost, Gradient Boosting, and Logistic Regression; for SMOTE, the top three models are Gaussian Naive Bayes (GNB), K-Nearest Neighbors (KNN), and AdaBoost.

Table 2: Top 3 models for each method.

Method	Model	Macro Recall
/	Dummy Classifier (baseline)	0.3333
Original Data	**Logistic Regression**	**0.5137**
	Neural Network	0.4950
	Support Vector Machine (rbf)	0.4878
Undersampled Data	**AdaBoost**	**0.5067**
	Gradient Boosting	0.4986
	Logistic Regression	0.4979
Oversampled Data	**Adaboost**	**0.5087**
	Gradient Boosting	0.5078
	Logistic Regression	0.4954
SMOTE	**Gaussian Naive Bayes**	**0.4342**
	K-Nearest Neighbors	0.4193
	Neural Network	0.3864

After parameter tuning, the model with the highest macro recall on the original dataset is Logistic Regression (when 'penalty' is set to 'none'); for undersampling method, the best model is Gradient Boosting (when 'n estimators' is set to 150); for oversampling method, the best model is AdaBoost (when 'n estimators' is set to 150); for SMOTE method, the best model is Gaussian Naive Bayes with the default parameters.

Table 3: The best model for each method.

Method	Model	Macro Recall
Original Data	Logistic Regression	0.5139
Undersampled Data	Gradient Boosting	0.5083
Oversampled Data	Adaboost	0.5138
SMOTE	Gaussian Naive Bayes	0.4342

5.3 Results Analysis

From the outcomes above, it can be concluded that the best model is Logistic Regression with no penalty under the original data, which can reach 0.5139 of macro recall. 0.1806 has been improved compared with the macro recall of the baseline. We could not find studies that aimed to classify no diabetes, prediabetes, and diabetes using exactly the same dataset as ours. However, in Kuo

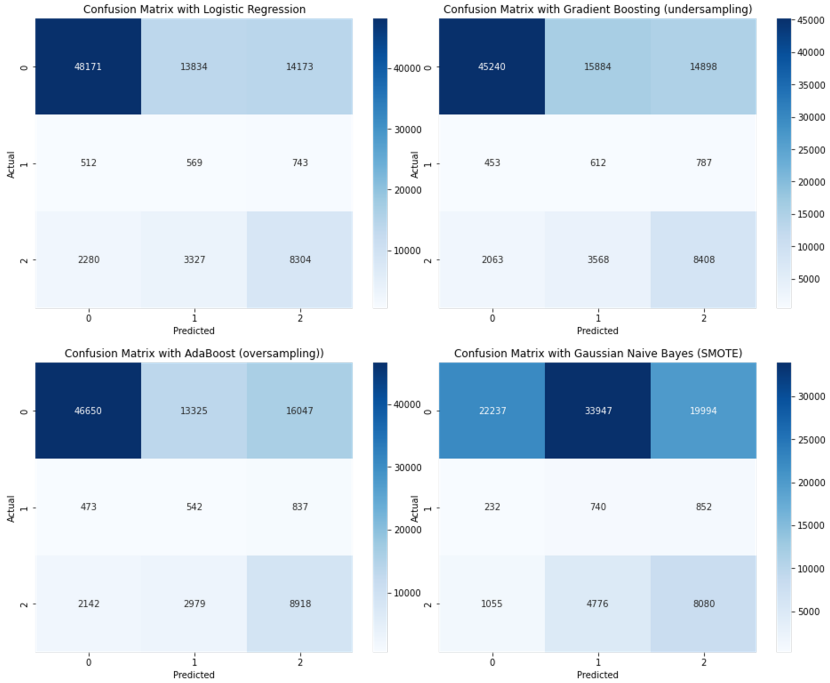

Fig. 5: Confusion Matrix of the Best Models.

et al. (2020)'s study, they reached the recall value of 0.991 using SVM model [23] for classifying the type of diabetes for those who are already diabetic. Most of the studies focus on binary classification. For example, in Xie et al.'s study, based on BRFSS 2014 data, the highest recall for binary classification is 0.5161 [8], which is similar to the highest macro recall of the multi-classification in our work.

6 Discussion

It can be seen from the outcomes that the performance of Logistic Regression is not worse than other models. Instead, its stability and robustness in handling both the original dataset and different sampling methods highlight its reliability and efficiency in handling classification tasks, showcasing its competitive edge in real-world applications. Joshi and Dhakal (2021) also proved that Logistic Regression can perform similarly to ensemble-based machine learning models [22].

Apart from the Logistic Regression model, the Neural Network was expected to perform well in a classification task. However, during the training process, the NN model's performance steadily improved on the training dataset (the recall of the training set was increasing), while failed to exhibit a similar enhancement on the test set (the recall of the test set stayed almost the same during training), indicating low generalized ability on the test set. Therefore, we supposed that the data was overfitting on the training set during the NN model training process. A possible solution is to create a validation set from the training set and stop the model training early if convergence in the validation performance is observed.

During the implementation of our work, we also noticed the importance of assigning and adjusting the class weight during training such that the classifier is trained towards classifying hard problems which in our case is to accurately predict the pre-diabetes and diabetes. Lin et al. (2018) proposed that a scaling factor $(1-p_t)^\gamma$ can be added to the standard cross entropy criterion $FL(p_t) = -(1-p_t)^\gamma log p_t$ to down-weight the contribution of easy examples during training and rapidly focus the model on hard examples [21]. Though originally for object detection, the focal loss could be our future direction for neural network training for the diabetes dataset.

In addition, as mentioned during the exploratory data analysis stage, the classifiers tend to confuse the pre-diabetes and diabetes status as evidenced by the confusion matrices (Fig. 5.). Adopting the idea of ensemble learning that takes advantage of more than one classifier, the classification performance might be further improved if an additional classifier that aims to separate pre-diabetes and diabetes is trained. Based on the outputs of the initial classifier, the additional classifier will re-classify instances that are predicted to be pre-diabetes or diabetes by the first classifier. Taking the classification results from both classifiers, the pre-diabetes and diabetes status are expected to be better distinguished.

7 Conclusion

In this work, we focused on constructing the best multi-class classification model with an appropriate sampling method to predict if a subject will contract diabetes or have a risk for pre-diabetes, which can be used for mass screening. To address the class imbalance inherent in the original dataset, we employed a range of resampling techniques, including undersampling, oversampling, and SMOTE. Additionally, we evaluated the performance of various classification models, encompassing Logistic Regression, Gaussian Naive Bayes, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, K-Nearest Neighbors, and Support Vector Machine, in pursuit of achieving the highest macro recall. Following hyperparameter tuning for the top 3 classifiers, the most effective combination is applying Logistic Regression with no penalty to the original data, which has the highest macro recall with 0.5139. This substantial improvement over the baseline macro recall of 0.3333 underscores the model's capacity to significantly enhance the identification of individuals at risk. Consequently, our study contributes to the advancement of predictive healthcare analytics, offering a valuable tool for population-wide diabetes screening.

References

1. Aune, D., Norat, T., Leitzmann, M. et al. Physical activity and the risk of type 2 diabetes: a systematic review and dose–response meta-analysis. Eur J Epidemiol 30, 529–542 (2015). https://doi.org/10.1007/s10654-015-0056-z
2. Cerf ME. Beta cell dysfunction and insulin resistance. Front Endocrinol (Lausanne). 2013 Mar 27;4:37. doi: 10.3389/fendo.2013.00037
3. Drinkwater, JJ, Davis, WA, Davis, TME. A systematic review of risk factors for cataract in type 2 diabetes. Diabetes Metab Res Rev. 2019; 35:e3073. https://doi.org/10.1002/dmrr.3073
4. Diniz Pereira, J, Gomes Fraga, V, Morais Santos, AL, Carvalho, MG, Caramelli, P, Braga Gomes, K. Alzheimer's disease and type 2 diabetes mellitus: A systematic review of proteomic studies. J Neurochem. 2021; 156: 753–776. https://doi.org/10.1111/jnc.15166
5. Li Y, Schoufour J, Wang DD, Dhana K, Pan A, Liu X, Song M, Liu G, Shin HJ, Sun Q, Al-Shaar L, Wang M, Rimm EB, Hertzmark E, Stampfer MJ, Willett WC, Franco OH, Hu FB. Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: prospective cohort study. BMJ. 2020 Jan 8;368:l6669. doi: 10.1136/bmj.l6669
6. Zurita-Cruz JN, Manuel-Apolinar L, Arellano-Flores ML, Gutierrez-Gonzalez A, Najera-Ahumada AG, Cisneros-González N. Health and quality of life outcomes impairment of quality of life in type 2 diabetes mellitus: a cross-sectional study. Health Qual Life Outcomes. 2018 May 15;16(1):94. doi: 10.1186/s12955-018-0906-y
7. Han H, Cao Y, Feng C, Zheng Y, Dhana K, Zhu S, Shang C, Yuan C, Zong G. Association of a Healthy Lifestyle With All-Cause and Cause-Specific Mortality Among Individuals With Type 2 Diabetes: A Prospective Study in UK Biobank. Diabetes Care. 2022 Feb 1;45(2):319-329. doi: 10.2337/dc21-1512

8. Carmichael J, Fadavi H, Ishibashi F, Shore AC, Tavakoli M. Advances in Screening, Early Diagnosis and Accurate Staging of Diabetic Neuropathy. Front Endocrinol (Lausanne). 2021 May 26;12:671257. doi: 10.3389/fendo.2021.671257
9. Thipsawat S. Early detection of diabetic nephropathy in patient with type 2 diabetes mellitus: A review of the literature. Diab Vasc Dis Res. 2021 Nov-Dec;18(6):14791641211058856. doi: 10.1177/14791641211058856
10. Hosler, Akiko S. PhD; Berberian, Elizabeth L. MPH; Spence, Maureen M. MS, RD; Hoffman, David P. MEd. Outcome and Cost of a Statewide Diabetes Screening and Awareness Initiative in New York. Journal of Public Health Management and Practice 11(1):p 59-64, January 2005.
11. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2014 Nov 15;13:8-17. doi: 10.1016/j.2014.11.005
12. Field M, Hardcastle N, Jameson M, Aherne N, Holloway L. Machine learning applications in radiation oncology. Phys Imaging Radiat Oncol. 2021 Jun 24;19:13-24. doi: 10.1016/j..2021.05.007
13. Yang F, Zhang Q, Ji X, Zhang Y, Li W, Peng S, Xue F. Machine Learning Applications in Drug Repurposing. Interdiscip Sci. 2022 Mar;14(1):15-21. doi: 10.1007/s12539-021-00487-8
14. Dinh, A., Miertschin, S., Young, A. et al. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 19, 211 (2019). https://doi.org/10.1186/s12911-019-0918-5
15. Oikonomou EK, Khera R. Machine learning in precision diabetes care and cardiovascular risk prediction. Cardiovasc Diabetol. 2023 Sep 25;22(1):259. doi: 10.1186/s12933-023-01985-3
16. Chen C, Liaw A, Brieman L (2004) Using random forest to learn imbalanced data: Technical Report No. 666. University of California, Berkley. Using Random Forest to Learn Imbalanced Data.
17. Sadeghi S, Khalili D, Ramezankhani A, Mansournia MA, Parsaeian M. Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods. BMC Med Inform Decis Mak. 2022 Feb 10;22(1):36. doi: 10.1186/s12911-022-01775-z
18. Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev Chronic Dis 2019;16:190109. doi: http://dx.doi.org/10.5888/pcd16.190109
19. Dagliati A, Marini S, Sacchi L, et al. Machine Learning Methods to Predict Diabetes Complications. Journal of Diabetes Science and Technology. 2018;12(2):295-302. doi:10.1177/1932296817706375
20. Sadeghi S., Khalili D., Ramezankhani A. et al. Diabetes mellitus risk prediction in the presence of class imbalance using flexible machine learning methods. BMC Med Inform Decis Mak 22, 36 (2022). https://doi.org/10.1186/s12911-022-01775-z
21. Lin T., Goyal P., Girshick R. et al. Focal Loss for Dense Object Detection. Accessed December 12, 2023. doi: https://doi.org/10.48550/arXiv.1708.02002
22. Joshi R, Dhakal C. Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. International Journal of Environmental Research and Public Health. 2021 Jul 9;18(14):7346. doi: 10.3390/ijerph18147346.
23. Kuo K, Talley P, Kao Y, Huang C. A multi-class classification model for supporting the diagnosis of type II diabetes mellitus. PeerJ. 2020 Sep 10;8:e9920. doi: 10.7717/peerj.9920.
24. *Sklearn.linear_model.logisticregression.* scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

25. Teboul, A. (2021, November 8). *Diabetes health indicators dataset*. Kaggle. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

8 Appendix

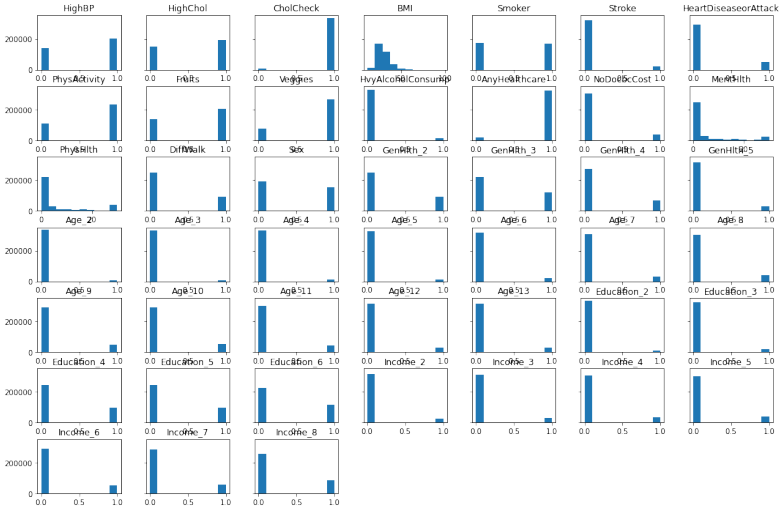

Fig. 6: Feature Distribution in the Oversampled Data

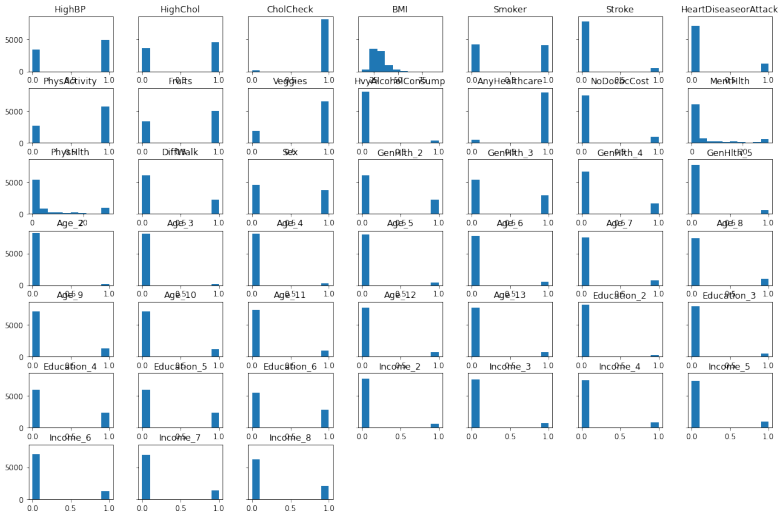

Fig. 7: Feature Distribution in the Undersampled Data

Fig. 8: Feature Distribution in the SMOTE Data