Diabetes Determinants

Connor Donahue
College of Computing
Michigan Technological University
Houghton, MI, USA
cdonahue@mtu.edu

***Abstract*—This report utilizes a diabetes health indicator dataset, composed of survey data including health, lifestyle, and demographic information collected from 70,692 individuals. Individuals are classified according to two categories: diabetic or prediabetic, and normal. The purpose of this study was to create an accurate classification model intended to predict whether an individual may be diabetic or prediabetic. To achieve this goal, a tree ensemble model was constructed using the xgboost algorithm and hyperparameter tuning on the number of estimators and the maximum tree depth of the model. This analysis resulted in a model that scored a moderate ROC AUC of 0.83 on the held-out test data. The xgboost model outperformed other common classification models including a decision tree, k-nearest neighbor classifier, and bernoulli naive bayes classifier. While it performed well relative to other models and demonstrated moderate discriminative ability, our xgboost classifier failed to outperform models from similar studies. This is likely due to the nonspecific nature of our dataset, which was collected as a general health survey and not focused solely on diabetes. Analyzing the relative importance of each feature in our model suggests that lifestyle data such as physical activity, drinking habits, and smoking habits are not as useful for predicting diabetes diagnosis as traditional medical data.**

***Index Terms*—diabetes, classification, machine learning, healthcare**

I. Introduction

Diabetes is characterized by the inability to regulate blood glucose levels [1]. Type I diabetes entails an inability to produce insulin, a molecule that promotes glucose uptake. Type II diabetes involves improper use of insulin and/or decreased insulin production. Diabetes can significantly decrease one's quality of life and life span. Once viewed as a rare disease, diabetes has become a global epidemic [2]. Cultural shifts resulting in decreased physical activity and the introduction of highly processed, sugary food into most of the world's diets are often blamed for the uptick in diabetes. In order to avoid contracting the disease, a lifestyle change is often required. Having a model that reliably predicts a diabetes diagnosis from basic health data including BMI and cholesterol, lifestyle data, and demographic data can aid individuals in understanding the risks their current lifestyle entails. Such a model can also be used to determine whether or not certain lifestyle changes need to be made. The goal of this report is to construct an accurate classification model that can be used to predict the likelihood of an individual having diabetes based on their lifestyle choices, basic health data, and demographic information.

II. Related Works

Predicting and diagnosing common diseases is a frequent subject of machine learning research. As such, there have been many studies directed towards diagnosing diabetes [3-6]. The most prominent research in this field has generally found that tree ensemble models, including random forest and two-class boosted decision tree, perform the best [3-6]. For example, Zou et al. achieved an accuracy of 0.8084 training a random forest model on data from physical examination records from Luzhou, China. The training dataset contained measurements of the age, pulse rate, breathing rate, left systolic pressure, right systolic pressure, left diastolic pressure, right diastolic pressure, height, weight, physique index, fasting glucose, waistline, low density lipoprotein, high density lipoprotein, and diabetic status of 68,994 individuals [3]. In a similar study, Soni and Varma found that their random forest model achieved an accuracy of 0.77, outperforming k-nearest neighbor, logistic regression, decision tree, support vector machine, and gradient boosting [4]. Soni and Varma utilized data from 768 female individuals of Pima Indian heritage. The dataset contained features tracking number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, body mass index (BMI), diabetes pedigree function, and age of each individual. Most impressively, Chou et al. trained a two-class boosted decision tree with an area under the curve score of 0.99 [5]. The data used to train and test the model were collected from 15,000 women aged between 20 and 80 during outpatient examinations at a Taipei Municipal medical center. The data features correspond to each individual's number of pregnancies, plasma glucose level, diastolic blood pressure, sebum thickness, insulin level, BMI, diabetes pedigree function, and age.

While these previous studies have demonstrated that powerful classifiers can be trained on data gathered from blood tests and other medical procedures, models trained on simple survey data have exhibited equally strong performance. For example, Gaine and Malik achieved an area under the curve of 0.99 with a bagged decision tree model trained on survey data with features including age, sex, smoking status, drinking status, frequency of thirst, frequency of urination, height, weight, and fatigue status [6]. The dataset used contains 1,939 records that were collected through online and in-person methods from subjects in Jammu and Kashmir. Both genders and a variety of social and economic demographics were represented in the

dataset. Despite relying on simple training data, this strong result suggests that well-designed surveys can be effective diagnostic tools. The features in Gaine and Malik's data are a prime example of thoughtful survey design because frequent urination, constant fatigue, and excessive thirst have been identified as the most frequently reported symptoms of both type I and type II diabetes [7].

III. Data Description

The dataset used in this study, titled "Diabetes Health Indicators Dataset," comes from Kaggle [8]. The data was collected by the CDC-sponsored Behavioral Risk Factor Surveillance System (BRFSS), which is an annual health-related telephone survey collecting responses from over 400,000 Americans each year. This study uses a cleaned set of 70,692 responses from the 2015 BRFSS survey that does not include any individuals with missing values. The reduced number of samples is due to the original dataset being resampled in order to achieve a 50/50 split between the positive and negative target class. This ensures that the model will not be biased towards predicting the more frequently occurring class over the less common one. The target class is represented by a binary indicator variable that equals 1 for individuals with diabetes or prediabetes and 0 for those without either condition. The remaining 21 features of the dataset, to be used for predicting the target class, are listed below:

- "HighBP": 0 = no high blood pressure, 1 = high blood pressure
- "HighChol": 0 = no high cholesterol, 1 = high cholesterol
- "CholCheck": 0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years
- "BMI": Body Mass Index (numeric value)
- "Smoker": 0 = never smoked or smoked less than 100 cigarettes, 1 = smoked at least 100 cigarettes
- "Stroke": 0 = no stroke history, 1 = stroke history
- "HeartDiseaseorAttack": 0 = no coronary heart disease or myocardial infarction, 1 = yes
- "PhysActivity": 0 = no physical activity in past 30 days, 1 = yes physical activity
- "Fruits": 0 = less than one fruit per day, 1 = one or more fruits per day
- "Veggies": 0 = less than one vegetable per day, 1 = one or more vegetables per day
- "HvyAlcoholConsump": 0 = not heavy drinker, 1 = heavy drinker (men $>$ 14 drinks/week, women $>$ 7 drinks/week)
- "AnyHealthcare": 0 = no healthcare coverage, 1 = has healthcare coverage
- "NoDocbcCost": 0 = did not avoid doctor due to cost, 1 = avoided doctor due to cost
- "GenHlth": General health rating (1 = excellent to 5 = poor)
- "MentHlth": Number of days mental health was not good in past 30 days (numeric)
- "PhysHlth": Number of days physical health was not good in past 30 days (numeric)
- "DiffWalk": 0 = no difficulty walking or climbing stairs, 1 = yes difficulty
- "Sex": 0 = female, 1 = male
- "Age": Age category (numeric code from 1 = 18–24 to 13 = 80+)
- "Education": Education level (numeric code from 1 = never attended school to 6 = college graduate)
- "Income": Income category (numeric code from 1 = $<$ \$10,000 to 8 = $\geq$ \$75,000)

IV. Methods

A. Pre-processing

The data pre-processing step was relatively simple because the data is already cleaned and contains balanced class sizes for the target variable. In order to remove training bias towards any particular feature, the non-binary features "BMI," "GenHlth," "MentHlth," "PhysHlth," "Age," "Education," and "Income" were scaled onto the range (0,1) using min-max scaling. The data was also split into two groups using random stratified sampling. 80% of samples were included in the training + validation set and the other 20% were used as the test set for model evaluation. A 2-component principal component analysis (PCA) was performed in order to explore any significant clustering between diabetic and nondiabetic individuals. The results, plotted in Figure 1, demonstrate weak clustering among members of the same class, suggesting that dimensionality reduction is not advisable for this dataset.

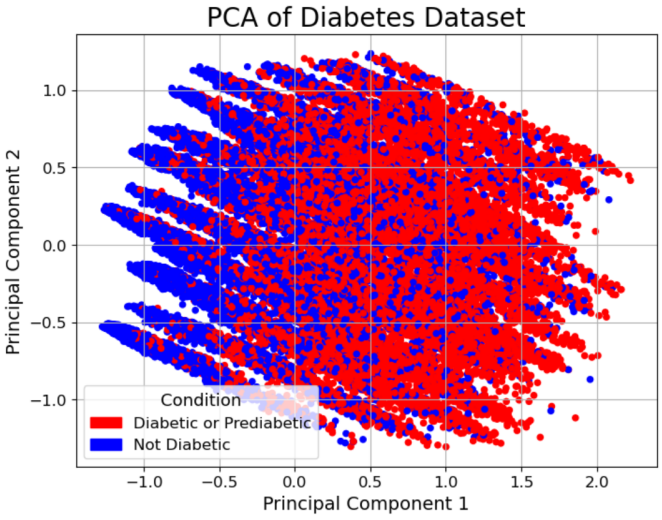

Fig. 1. Scatter plot of training + validation data using two principal components.

B. Model Training and Cross-validation

In light of the successful ensemble methods presented in the literature, an XGBoost model was fitted to the data. XGBoost is one of the flagship gradient boosting tree ensemble algorithms [9]. Gradient boosting operates by iteratively fitting small decision trees to the residual errors of the previous trees in the model. The model's learning rate was set to 0.1 and

hyperparameter tuning on the number of estimators (number of trees) and maximum depth of each estimator was conducted. Each combination of hyperparameters was assessed using 5-fold cross-validation in order to avoid selecting a combination that might achieve the highest score by chance. The evaluation method of choice was the receiver operating characteristic area under the curve (ROC AUC), which indicates the model's ability to make correct predictions by comparing the true positive rate with the false positive rate. Finally, the accuracy, precision, and recall were calculated for the best performing model.

In order to benchmark the model's performance, additional machine learning models were trained and evaluated. These include a decision tree (DT) classifier, naive bayes (NB) classifier, and k-nearest neighbors (KNN) classifier. The DT was trained using the default hyperparameters of the DecesionTreeClassifier() model available in the Scikit-learn library. Similarly, the NB classifier was also trained using the BernoulliNB() Scikit-learn model with default hyperparamters. Scikit-learn's KNNClassifier() was used for the KNN model, with the k-value set to 5. The ROC AUC scores were assessed for each of the models and compared with that of the xgboost model.

Following model comparisons, further investigation into the xgboost model was conducted. Each feature was ranked according to its importance. Importance was evaluated using the gain metric, which corresponds to the reduction of the log loss that the feature contributed to the model.

V. Experiment and Results

The machine learning models achieved the following ROC AUC scores: xgboost, 0.83; DT, 0.65; KNN, 0.76; NB, 0.78 (see Figure 2). All combinations of hyperparameters for which the xgboost model was evaluated achieved cross-validation ROC AUC scores ranging from 0.828 to just over 0.830, suggesting that hyperparameter tuning had little impact on the model's performance. Additionally, there were no significant trends in performance across the range of different hyperparameter values that were evaluated (see Figure 3). Therefore, despite the lack of hyperparameter tuning for the other classifiers, it is fair to conclude that the xgboost model performed the best of all models evaluated.

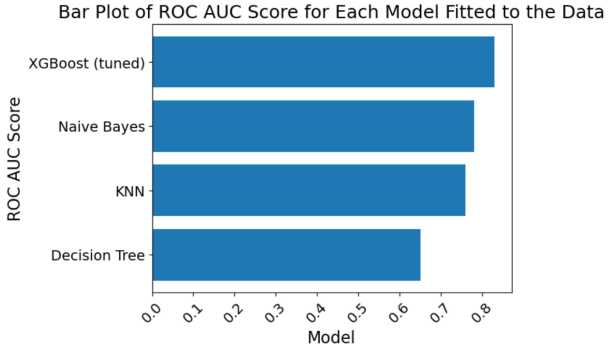

Fig. 2. Bar plot of ROC AUC scores for each model.

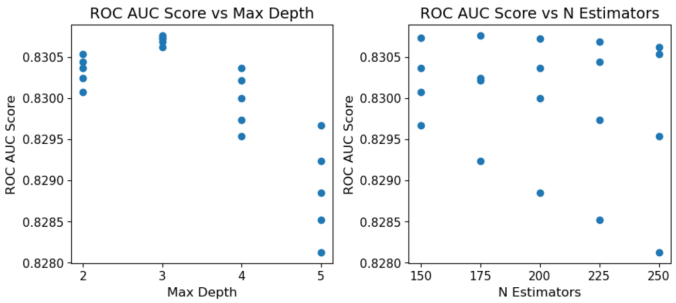

Fig. 3. Cross-validation ROC AUC scores plotted against different hyperparameter values.

In addition to a moderate-to-strong ROC AUC score of 0.83, the xgboost model achieved an accuracy of 0.75, precision of 0.73, and recall of 0.80. A plot of the data features, ranked by importance, is included in Figure 4. Evidently, the "HighBP" (gain of 794.00), "GenHlth" (gain of 296.67) and "HighChol" (gain of 156.55) features are most influential. This suggests that the additional lifestyle data such as "Smoker" (gain of 5.27), "PhysActivity" (gain of 4.76), "Fruits" (gain of 4.09), and "Veggies" (gain of 2.49) have little importance in classifying individuals by their diabetes diagnosis.

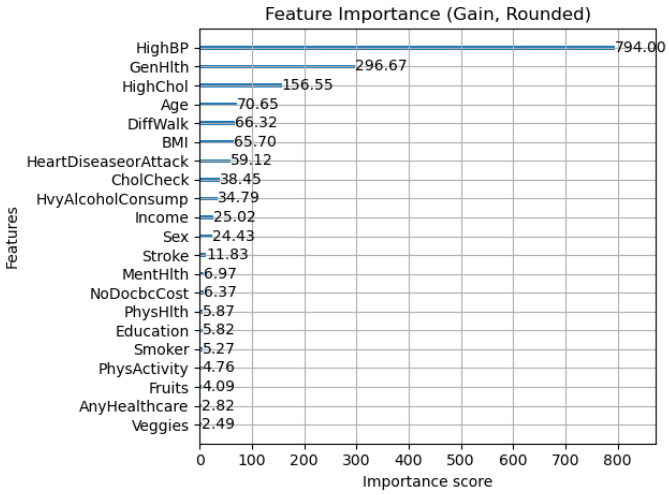

Fig. 4. Bar Plot of features ranked by importance (gain).

VI. Conclusion

Similar to the literature reviewed in section II of this study, we found that a tree ensemble model, namely xgboost, was the best classifier for the diabetes prediction task. Our strongest model was able to achieve moderate levels of predictive power with a ROC AUC of 0.83 and accuracy of 0.75, but it was unable to outperform the results from the related literature, which achieved accuracies ranging from 77-99% [3, 6]. The hyperparameter tuning that was performed had an insignificant impact on the model's performance, suggesting that the training dataset may be the cause of the model's relatively weak performance.

Both the manner of data collection and the data itself are atypical in comparison to the datasets on which other models were trained. For example, most of the available literature uses records from medical examinations [3-5], which ensures some level of measurement accuracy and standard procedure in data collection. As such, these other datasets also included more numerical features, such as blood pressure and insulin level, that are highly informative of one's diabetic status yet can only be collected via medical procedures. Unfortunately, the dataset used in this study relied on self-reported metrics that were simplified into binary indicators. For example, the "HighBP" feature in our dataset equals 0 for no high blood pressure and 1 for high blood pressure, reducing the range of values that can be used for prediction.

In addition to issues in data collection and measurement rigor, our dataset faces the simple disadvantage of having not been designed for the purpose of diabetic classification. To illustrate this idea, the study conducted by Gaine and Malik demonstrates that simple survey-style data can be used to effectively predict diabetic status when the data is tailored to the task [6]. Gaine and Malik found that the two most influential features in their model were simple variables indicating how often the individual peed at night and whether or not they felt fatigued. Had these been captured by the BRFSS survey questions, our model may have achieved drastically stronger performance. This was not the case, however, because the BRFSS survey from which our data originates is a general health survey that does not focus on diabetes [8].

In order to improve upon the model created in this study, further hyperparameter tuning could be performed; model performance did not improve significantly across different values for the number of estimators and max depth of each estimator, but the xgboost algorithm possesses various other hyperparameters that can be altered. A future study, for instance, might find that decreasing the model's learning rate would yield better predictive power. However, recording new features would likely improve the model's performance significantly more than hyperparameter tuning; as previously mentioned, adding additional questions regarding fatigue and urination frequency to the survey would be highly informative of an individual's diabetic status.

Ultimately, this study reaffirmed the superiority of tree ensemble models over other simple machine-learning models and demonstrated that data from simple survey questions can achieve moderate performance in the diabetes classification task. Model performance was not strong enough to be applied in any clinical scenarios, but it does suggest that more targeted survey questions may provide the basis for diabetes diagnosis without requiring expensive and sometimes inaccessible medical examinations. At the very least, predictive models may be used to provide insights that can guide patients in the decision of whether or not to schedule such examinations.

REFERENCES

[1] G. Roglic, "WHO Global report on diabetes: A summary," *International Journal of Noncommunicable Diseases*, vol. 1, no. 1, pp. 3–8, 2016.

[2] P. Z. Zimmet *et al.*, "Diabetes: a 21st century challenge," *The Lancet Diabetes and Endocrinology*, vol. 2, no. 1, pp. 56–64, 2014.

[3] Q. Zou *et al.*, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, Art. no. 515, 2018.

[4] M. Soni and S. Varma, "Diabetes prediction using machine learning techniques," *International Journal of Engineering Research and Technology*, vol. 9, no. 9, pp. 921–925, 2020.

[5] C.-Y. Chou, D.-Y. Hsu, and C.-H. Chou, "Predicting the onset of diabetes with machine learning methods," *Journal of Personalized Medicine*, vol. 13, no. 3, Art. no. 406, 2023.

[6] S. M. Ganie and M. B. Malik, "An ensemble machine learning approach for predicting type-II diabetes mellitus based on lifestyle indicators," *Healthcare Analytics*, vol. 2, Art. no. 100092, 2022.

[7] N. G. Clark *et al.*, "Symptoms of diabetes and their association with the risk and presence of diabetes: findings from the Study to Help Improve Early evaluation and management of risk factors Leading to Diabetes (SHIELD)," *Diabetes Care*, vol. 30, no. 11, pp. 2868–2873, 2007.

[8] A. Teboul, 2021, "Diabetes Health Indicators Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset.

[9] C. Tianqi and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.