

Developing Classification Models to Predict Diabetes

Lillian Mueller lmuelle1@umd.edu

Abstract— Understanding the role predictive modeling could have in the realm of diabetes prevention and maintenance holds immense potential in advancing personalized healthcare and facilitating early intervention. Leveraging a dataset consisting of diverse health indicators sourced from Kaggle, the efficacy of multiple models in identifying individuals at risk was investigated. This study examines various classification models including decision tree, logistic regression, k-nearest neighbor, naïve Bayes classifier, and linear discriminant analysis. Overall, the Gaussian naïve Bayes classification model exhibited the highest precision, emphasizing its capability in identifying individuals with prediabetes. However, the dataset’s imbalance significantly impacted the model’s overall accuracy, underscoring the need to address this issue. Insights into the comparative effectiveness of classification models for diabetes prediction can be expanded to other preventable condition given specific health indicators.

I. INTRODUCTION

Diabetes is a prevalent metabolic condition that affect millions of people globally. In the United States, diabetes is the eighth leading cause of death [1]. The breadth of these conditions propels research to develop predictive methods to enable timely intervention and management. Understanding these conditions and building the capacity to predict them has great potential to improve individual health outcomes and societal outlook.

Characterized by elevated blood sugar levels, diabetes encompasses type 1, type 2, and gestational diabetes. Type 1 diabetes occurs when the body is unable to make insulin, which is a hormone that regulates glucose levels in the blood. Type 2 diabetes is the most common form of diabetes; an individual with this type cannot properly utilize insulin. Gestational diabetes can occur during pregnancy but increases the mother’s chance of getting type 2 diabetes later in life [1]. Type 2 is the most common and can develop at any age, it can also be prevented. Prediabetes is a stepping stone that could be used as an indicator to being at risk of type 2 diabetes. Having prediabetes means your blood glucose levels are higher than normal [1]. Being able to identify this condition is crucial for intervention strategies such as lifestyle modifications in the form of weight loss and exercise [2].

This investigation explored the predictive capabilities of various classification models in determining whether an individual has diabetes, prediabetes, or no diabetes at all. The models under scrutiny include the decision tree, logistic regression, k-nearest neighbor (KNN), naïve Bayes classifier, and linear discriminant analysis (LDA). Each model offers unique methodologies and algorithms that may exhibit varying efficacy in predicting diabetes or prediabetes based on health indicators.

The investigation utilizes a comprehensive dataset sourced from Kaggle [3]. This dataset consists of health indicators such as demographic information, physiological measures, and lifestyle attributes. These attributes include age, body mass index (BMI), blood pressure, insulin levels, blood pressure, heart conditions, diet, exercise, and more. By examining these models against the dataset, this study seeks to provide insights into the most effective predictive framework for identifying individuals at risk or already affected by these metabolic conditions.

Section II describes how Python was used to clean the data, develop the models, and compare the models’ performance. The final determination of the best fit model is in Section III. Section IV contains discussion about performance evaluation and future research.

II. METHODOLOGY

This investigation entails using various classes from the library `scikitlearn`. To develop the models: `GaussianNB` class from the `naive_bayes` module, the `DecisionTreeClassifier` class from the `tree` module, the `LogisticRegression` class from the `linear_model` module, the `KNeighbors` function from the `neighbors` module, and `LinearDiscriminantAnalysis` class from the `discriminant_analysis` module. Each was used to develop the naïve bayes classifier, decision tree model, the logistic regression model, KNN model, and the LDA respectively. To develop the confusion matrices and ROC (receiver operating characteristics) of each model, `sklearn`’s `metrics` module was used. Other Python libraries used along the way to help handle the data include `matplotlib` → `.pyplot`, `pandas`, and `numpy`.

A. Diabetes Dataset Preprocessing

The dataset in this investigation was taken from Kaggle [3]. The data was downloaded from the website as a CSV file, which is a file type where the data is separated by commas. The `pandas` library has methods that allows the analyst to directly read the data from a CSV file into a dataframe. Once the data was translated into the dataframe, each column was examined and cleaned to contain only numeric values and was scaled. The target classification column contained the distinction of whether an individual had diabetes, prediabetes, or no diabetes. After encoding these classifications, 0 represented no diabetes, 1 represented prediabetes, and 2 represented diabetes. All other columns in the dataframe

represented attributes that are used as features in the models to predict these classifications. The columns were all altered to only contain numerical values. The final values for each attribute are as follows:

- 1) High Blood Pressure (HighBP)
 - 0 = no
 - 1 = yes
- 2) High Cholesterol (HighChol)
 - 0 = no
 - 1 = yes
- 3) Cholesterol Check within the past 5 years (CholCheck)
 - 0 = no
 - 1 = yes
- 4) Body Mass Index (BMI)
- 5) Have you smoked 100 cigarettes (5 packs) over lifetime? (Smoker)
 - 0 = no
 - 1 = yes
- 6) Stroke
 - 0 = no
 - 1 = yes
- 7) Coronary Heart Disease or Myocardial Infarction (Heart-DiseaseorAttack)
 - 0 = no
 - 1 = yes
- 8) Physical Activity within past 30 days (not including job) (PhysActivity)
 - 0 = no
 - 1 = yes
- 9) Consumes Fruit ≥ 1 times a day (Fruits)
 - 0 = no
 - 1 = yes
- 10) Consumes Vegetables ≥ 1 times a day (Veggies)
 - 0 = no
 - 1 = yes
- 11) Heavy Drinker men >14 drinks, women >7 drinks (HvyAlcoholConsump)
 - 0 = no
 - 1 = yes
- 12) Health Care Coverage (AnyHealthcare)
 - 0 = no
 - 1 = yes
- 13) Couldn't see doctor due to cost in past 12 months (NoDocbcCost)
 - 0 = no
 - 1 = yes
- 14) General Health Rating (GenHlth)
 - 1 = excellent
 - 2 = very good
 - 3 = good
 - 4 = fair
 - 5 = poor
- 15) Number of Days Mental Health was poor within past 30 days (MentHlth)
 - 1-30
- 16) Number of days physical/mental health kept individual from usual activities (PhysHlth)
 - 1-30
- 17) Difficultly walking/climbing stairs (DiffWalk)
 - 0 = no
 - 1 = yes
- 18) Sex
 - 0 = female
 - 1 = male
- 19) Age
 - 1 = ages 18-24
 - 2 = ages 25-29
 - 3 = ages 30-34
 - 4 = ages 35-39
 - 5 = ages 40-44
 - 6 = ages 45-49
 - 7 = ages 50-54
 - 8 = ages 55-59
 - 9 = ages 60-64
 - 10 = ages 65-69
 - 11 = ages 70-74
 - 12 = ages 75-79
 - 13 = 80 and older
- 20) Education
 - 1 = never attended school, only kindergarden
 - 2 = grades 1-8, elementary
 - 3 = grades 9-11, some high school
 - 4 = grade 12 or GED (high school graduate)
 - 5 = College 1-3 years (Some college or technical school)
 - 6 = college ≥ 4 years (college graduate)
 - 9 = refused
- 21) Income
 - 1 \leq \$10,000
 - 2 \leq \$15,000
 - 3 \leq \$20,000
 - 4 \leq \$25,000
 - 5 \leq \$35,000
 - 6 \leq \$50,000
 - 7 \leq \$75,000
 - 8 \geq \$75,000

Since there were 21 total features in the dataset, to decrease complexity and potentially increase accuracy of the models, attributes not directly related with the individuals' physical health and subjective attributes were omitted such as education, income, health care coverage, general health rating, and estimated number of days of poor mental/physical health. Overall, the dataset breakdown revealed that 84.2% of the data represented no diabetes, 1.8% of the dataset is classified as having prediabetes, and the remaining 14% of the data had diabetes.

After the diabetes dataset was processed, it was split into a train group and test group using the `model_selection.train_test_split()` function, with 2/3 of the data as the training group and the remainder as the testing group. For this dataset, since the classifications are not even, the test sets were stratified with respect to the diabetes classifications to attempt to split dataset with uniform proportions. Each model described in the following section was trained and tested using these training and testing data subsets.

B. Model Development

Next, the models were developed and each model was tuned to the dataset. The first model to be trained was the decision tree model. To tune this model, the entropy and Gini index criterion were tested. Due to the large number of features, the decision trees were also constrained by varying maximum depths. Following the procedures outlined in a previous study titled “Investigating Decision Trees”, the best fitting decision tree model was found to be utilizing the Gini index as the purity measurement and no constraints on the maximum depth of the tree [4].

Since this is a multiclass dataset, `SKLearn` has various solver algorithms to use in the optimization problem used in the logistic regression model. Those include Newton’s method, the Stochastic average gradient (SAG), a variant of the SAG the Stochastic average gradient descent (SAGA), and finally the Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm. For this application, only SAG and SAGA were investigated due to their ability to deal with large, multiclass datasets [5]. Both algorithms were tested with the dataset using no penalty and the l2 penalty. These penalties are further described in a previous study titled, “Iris Classification Using Logistic Regression” [6]. Upon running these variations of the logistic regression model, the model using SAGA solver without any penalties proved to have the greatest accuracy.

To develop the k-nearest neighbor (KNN) model, multiple parameters were tested to tune the model to the diabetes dataset. These parameters included modifying the number of neighbors, the distance metrics used to determine similarity, and simply testing the model using uniform weights. The distance metrics in question were Euclidean, Manhattan, and cosine. Procedures to develop these methods are defined in “Evaluating the Performance of K-Nearest Neighbors Classification” [7]. The optimal parameters proved to be $k = 25$ using the Euclidean distance metric.

For this dataset, two versions of the naïve bayes classifier could be a fit: the Gaussian naïve Bayes and the Bernoulli naïve Bayes. The Gaussian version assumes the feature values follows a Gaussian distribution whereas the Bernoulli naïve Bayes can be used when the features are boolean vectors [8]. The Gaussian naïve Bayes classifier is run using the procedure in the previous study, “Comparing Classifications Models Against the Naïve Bayes Classifier and Linear Discriminant Analysis Model” [9]. To develop the Bernoulli model, the `naive_bayes.BernoulliNB()` class was used.

Finally, the linear discriminant analysis (LDA) was developed using the methodology described in a previous study [9]. For this application, the default solver, the singular value decomposition, was used due to its ability to handle large numbers of features.

C. Classification Model Evaluation and Comparison

After determining the best version of each model for the diabetes dataset, they were evaluated based on performance and compared. To start, a confusion matrix was created for each model to better visualize how each model performed. The process by which to develop the confusion matrices and its visualization is described in “Evaluating Classifications Models using Confusion Matrices” [10]. These matrices enable the analyst to visually evaluate the models’ performance and derive trade-offs between true positives and false positives for each individual classification.

To compare and rank the models, a ROC curve was created for each model to find the models AUC (area under the curve). This value is a scalar value that allows the user to evaluate the models’ performance with respect to false positives versus true positives and provides an easy way for the analyst to rank the classifiers. The methodology to develop the curves for each multiclass model is similar to the methodology explained in previous reports [9], [11].

Finally, to further validate and compare these models, each model was evaluated using cross-validation. Using 10 folds and the methodology found in a previous investigation, the mean accuracy and the standard deviation of all the accuracies from each fold were found and compared [12]. Additionally, a similar method was used to perform cross validation to find the general precision score when classifying prediabetes. The `cross_val_score()` function was used here as well, expect rather than specifying the scoring as ‘accuracy’ \rightarrow , the scoring was based on ‘precision’. To isolate this classification in a multiclass dataset, the target array was binarized to highlight only prediabetes as a positive classification.

III. RESULTS

A. Model Accuracy and Precision

To reiterate, the best fit variation of the decision tree, logistic regression, and KNN models found for the diabetes dataset are as follows:

- Decision Tree utilizing the Gini Index impurity measurement with no maximum tree depth
- Logistic Regression utilizing the SAGA algorithm with no penalty
- K-Nearest Neighbor where $K = 25$ and the Euclidean distance is used

Each specialized model was chosen based on accuracy derived from cross-validation and the individual classification precision when compared to other variations of the same model. It is important to note that in many cases, the models

are unable to predict prediabetes. Upon observations, if a version of a model is able to predict some prediabetes classification at some accuracy, the overall model accuracy is reduced greatly. In this investigation, the model with the greatest accuracy and some competence at classifying prediabetes is chosen. If the model was unable to classify any prediabetes, the model with the greatest overall accuracy was chosen. The reason for this discrepancy is most likely due to the unbalanced dataset. Only about 2% of the dataset can be classified as prediabetes.

With regards to the naïve Bayes, both the Gaussian and Bernoulli versions of the naïve Bayes classifier had their own trade-offs. While the Gaussian naïve Bayes classifier was able to classify some prediabetes correctly, the overall accuracy was very poor in comparison. On the other hand, the Bernoulli naïve Bayes classifier had a very high overall accuracy but was unable to classify prediabetes. Table I shows the discrepancies between the models.

Model	Accuracy	Precision	Precision	Precision
		0	1	2
Decision Tree	0.785	0.833	0.014	0.361
Logistic Reg.	0.815	0.826	0.	0.521
KNN	0.814	0.826	0.	0.502
Gaussian NB	0.739	0.875	0.018	0.341
Bernoulli NB	0.797	0.833	0.	0.392
LDA	0.814	0.830	0.	0.497

TABLE I
MODEL ACCURACY AND PRECISION

According to these results, the logistic regression, KNN, and LDA models all have the greatest overall accuracy of 81%. The Gaussian naïve Bayes has the greatest precision when predicting no diabetes (88%) and prediabetes (1.8%), and the logistic regression model has the greatest precision when predicting diabetes at 52%.

B. Individual Classification Model Confusion Matrices

With the chosen classifiers to compare, the confusion matrix for each model was developed as seen in Figures 1 to 6. Confusion matrices enable the analyst to visually see the predictions made by each individual model and compare the predictions against the true classifications.

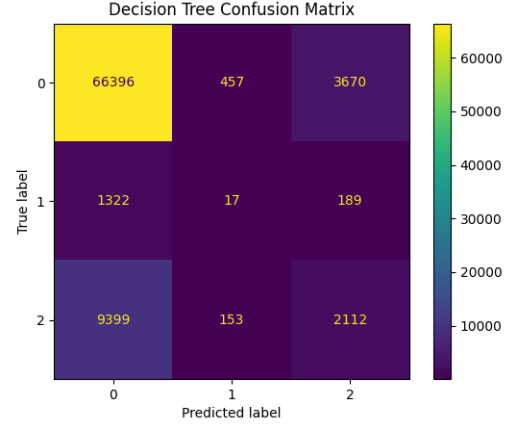


Fig. 1. Decision Tree Utilizing Gini Impurity Criterion Confusion Matrix

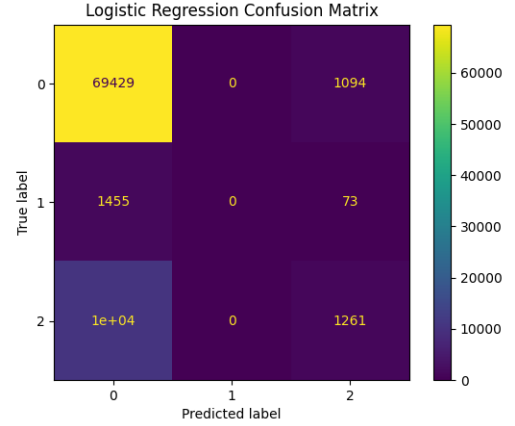


Fig. 2. Logistic Regression Utilizing SAGA Solver Confusion Matrix

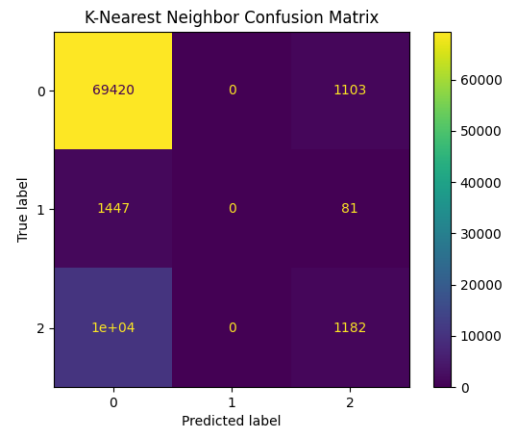


Fig. 3. KNN, K=25 Utilizing Euclidean Metric Confusion Matrix

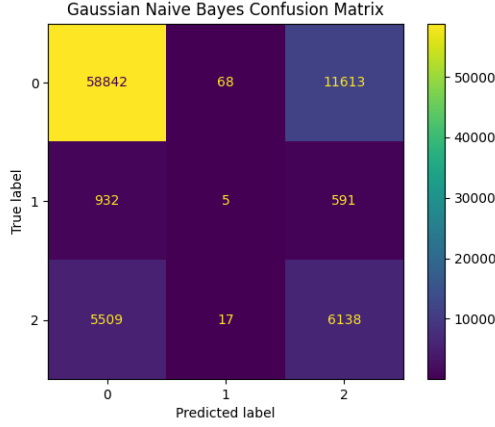


Fig. 4. Gaussian Naive Bayes Confusion Matrix

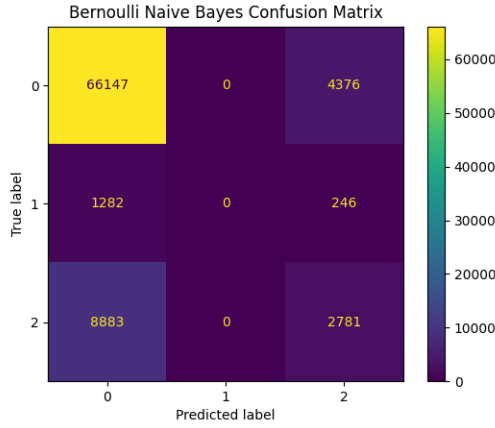


Fig. 5. Bernoulli Naive Bayes Confusion Matrix

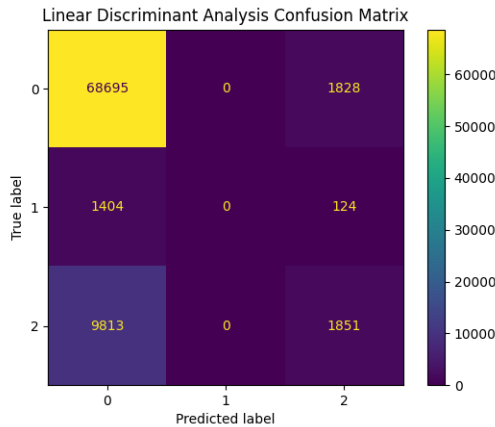


Fig. 6. LDA Confusion Matrix

From these results, the decision tree and Gaussian naïve Bayes are the only models capable to predicting prediabetes, however not very accurately. Objectively, these results are not very good. With the goal of being able to predict an individual's condition early enough for intervention, the low accuracy and precision when classifying prediabetes is concerning.

C. ROC Curves and AUC Rankings

To further compare the performance of these models, the receiver operating characteristics (ROC) curves were plotted for each model and the area under the curve (AUC) was calculated. The curves specific to classifying diabetes is in Figure 7, the curves specific to classifying prediabetes is shown in Figure 8, and the plots specific to classifying no diabetes is shown in Figure 9.

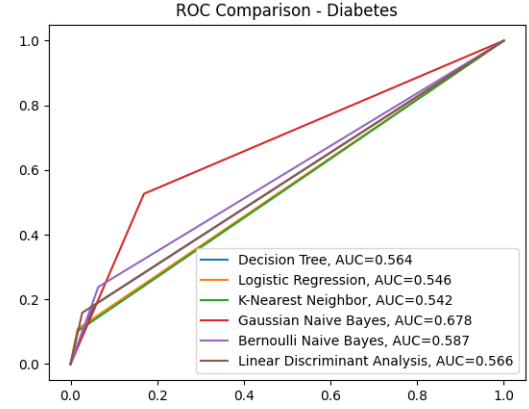


Fig. 7. ROC Curves for Diabetes Classification

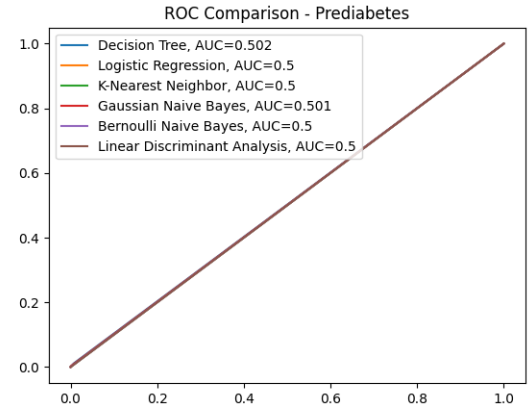


Fig. 8. ROC Curves for Prediabetes Classification

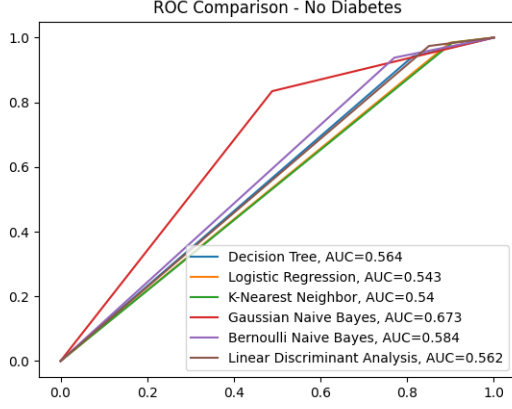


Fig. 9. ROC Curves for No Diabetes Classification

When ranking the models based on AUC based on each classification, the Gaussian Naïve Bayes proves to be the strongest performing model. All the rankings are shown in Table II.

Rank	No Diabetes	Prediabetes	Diabetes
1	Gaussian NB	Decision Tree	Gaussian NB
2	Bernoulli NB	Gaussian NB	Bernoulli NB
3	Decision Tree	-	Decision Tree
4	LDA	-	LDA
5	Logistic Reg.	-	KNN
6	KNN	-	Logistic Reg.

TABLE II
MODEL RANKINGS BASED ON AUC

D. Cross Validation

A 10-fold cross validation was performed on each model, both to find the general accuracy of the models and the general precision when classifying prediabetes. The results can be found in Table III and Table IV. The logistic regression and KNN models were found to be the most accurate with an accuracy of 84.3%. The computational time required to validate these models was also significantly longer than that of the other models. The model with the lowest accuracy was the Gaussian naïve Bayes with an accuracy of 77.5%. However, the Gaussian naïve Bayes had the greatest precision when predicting prediabetes; its precision rate was 29.9%.

Model	Mean Accuracy	Standard Dev.
Decision Tree	0.820	0.003
Logistic Reg.	0.843	0.005
KNN	0.843	0.001
Gaussian NB	0.775	0.009
Bernoulli NB	0.824	0.002
LDA	0.841	0.005

TABLE III
10-FOLD CROSS VALIDATION ACCURACY

Model	Mean Precision	Standard Dev.
Decision Tree	0.010	0.005
Logistic Reg.	0.000	0.000
KNN	0.000	0.000
Gaussian NB	0.298	0.399
Bernoulli NB	0.000	0.000
LDA	0.000	0.000

TABLE IV
10-FOLD CROSS VALIDATION PRECISION OF PREDICTING PREDIABETES

IV. DISCUSSION

Overall, the best classification model for this dataset was the Gaussian naïve Bayes classifier. This model was the only model capable of classifying the prediabetes class. However, if this were not the main objective of this investigation, based on these results, the analyst would have to decide what is more important: being able to classify prediabetes or the accuracy of making an overall classification. If the analyst values the capacity to predict prediabetes, the Gaussian naïve Bayes classifier would be the best fit. However, if the analyst values overall accuracy of the classification predictions, the logistic regression, KNN, or LDA models would be better. In general, these recommendations are only based on these models presented in the investigation. The low precision and accuracy may indicate a different model altogether should be examined for this dataset.

The discrepancy between the precision and accuracy of the predictions was unexpected. This investigation showed that a model's accuracy cannot be the only indicator of performance. Interestingly, the model with the lowest average accuracy was the Gaussian naïve Bayes. However, it showed the greatest precision when making individual classifications. For real-world applications, it is normal to have a high imbalance between classifications within a dataset. In this case, over 80% of the dataset indicated no diabetes and only 2% of the dataset made up the prediabetes category. The problem with using accuracy for a dataset with such imbalance is that accuracy treats all classes as equally important [13]. For this dataset and target classification problem, the precision is a more reliable indicator of performance as it measures how well the model predicts each individual classification. Using precision as the main performance indicator in this scenario is also very important as the cost of false positives is high; the model should therefore minimize false positives. This is crucial in healthcare scenarios (e.g. the classification of diabetes) where accurate identification of individuals at risk is very important.

In the future, addressing the dataset's imbalance would significantly enhance the models' accuracy. Some strategies to alleviate this issue include resampling the data, using a different model that can handle imbalanced datasets more efficiently, and collecting more data in general. If resampling the data, the analyst could either oversample the prediabetes data or undersample the diabetes and no diabetes classes. In addition to resampling, other models not presented in this investigation that can handle imbalanced datasets are models based on boosting or tree-based algorithms such as the AdaBoost classifier or the random forest classification model [14]. An-

other option that could enhance this classification problem is researching the most influential health indicators use in the diabetes prediction. For example, individuals over the age of 45 are more likely to be afflicted by diabetes [1]. Using this knowledge, more data can be collected from individuals in this age range, or the dataset could be reduced to only consist of individuals of the age 45 or over. In order to find a better fit model for classifying diabetes and prediabetes, the issue of dataset imbalance must be researched and remedied. Investigating the impact of imbalanced data could provide insights into the progression of diabetes and refine predictive models for long-term prognosis and intervention strategies.

REFERENCES

- [1] "About Prediabetes and Type 2 Diabetes — National Diabetes Prevention Program — CDC." Accessed: Nov. 23, 2023. [Online]. Available: <https://www.cdc.gov/diabetes/prevention/about-prediabetes.html>
- [2] J. P. Crandall et al., "The prevention of type 2 diabetes," *Nat. Clin. Pract. Endocrinol. Metab.*, vol. 4, no. 7, pp. 382-393, Jul. 2008, doi: 10.1038/ncpendmet0843.
- [3] "Diabetes Health Indicators Dataset." Accessed: Nov. 23, 2023. [Online]. Available: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>
- [4] L. Mueller and R. Hong, "Investigating Decision Trees".
- [5] ArnavR, "Scikit-learn solvers explained," Medium. Accessed: Nov. 25, 2023. [Online]. Available: <https://medium.com/@arnavr/scikit-learn-solvers-explained-780a17bc322d>
- [6] L. Mueller and R. Hong, "Iris Classification Using Logistic Regression".
- [7] L. Mueller and R. Hong, "Evaluating the Performance of K-Nearest Neighbors Classification".
- [8] S. Ray, "Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes Classifier," *Analytics Vidhya*. Accessed: Nov. 25, 2023. [Online]. Available: <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>
- [9] L. Mueller, "Comparing Classifications Models Against the Naive Bayes Classifier and Linear Discriminant Analysis Model".
- [10] L. Mueller, "Evaluating Classifications Models using Confusion Matrices".
- [11] L. Mueller, "Ranking Classification Models using Receiver Operating Characteristics".
- [12] L. Mueller and R. Hong, "Using K-Fold Cross Validation on Decision Tree and Logistic Regression Models to Classify Iris Species".
- [13] "Accuracy vs. precision vs. recall in machine learning: what's the difference?" Accessed: Nov. 26, 2023. [Online]. Available: <https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall>
- [14] R. Feki, "Imbalanced data: best practices," Medium. Accessed: Nov. 26, 2023. [Online]. Available: <https://rihab-feki.medium.com/imbalanced-data-best-practices-f3b6d0999f38>