Developing Classification Model to Predict Prediabetes and Diabetes

# Abstract

1. The paper is written as a journal article and NOT as a homework assignment.

2. The abstract provides an executive summary (each one in its own sentence).

a. Problem is summarized.

b. Basic methodology is provided (KNN, decision tree, etc.)

c. Results are summarized.

d. Why this is important and worth reading (discussion summary).

Understanding the intricacies of predictive modeling in the realm of diabetes and prediabetes holds immense potential in advancing personalized healthcare and facilitating early intervention.

# Introduction

3. The decision problem being solved is fully described.

4. How the decision will be used is provided.

5. Data set is described (here, not in other sections).

6. Literature review is provided. This is usually done in terms of what others have done before you in this field so it is clear that you are not repeating their efforts. For this class, feel free to include more references of other papers that talk about either the method you use or the problem you are solving. This could mean that you intent to try to get better results then them, or apply the same method used for one decision to another problem.

7. An overview of the rest of the paper (sections are defined) is provided.

Diabetes and prediabetes are prevalent metabolic conditions 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. Type 2 diabetes is the most common form of diabetes; an individual with this type cannot properly use insulin. Gestational diabetes can occur during pregnancy but increase the mother’s chance of getting type 2 diabetes later in life [1]. While 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 the likelihood of an individual having diabetes or prediabetes. The models under scrutiny include the decision tree, logistic regression, k-nearest neighbor, naive Bayes classifier, and linear discriminant analysis. 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, 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 \ref{methodology} describes 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 \ref{results}. Finally, the results and future research is discussed in Section \ref{discussion}.

# Methodology

8. Description of approach is provided.

9. Algorithm is described or a reference is provided.

10. Equations describing the approach are explained, not just presented, but explained.

11. Libraries and other software tools are defined.

12. Parameters of the libraries are discussed. Why did you use this value over that one, etc.

13. How results are compared or validated is provided. This needs to describe the method (take average, compute StDev, etc. and compare to ???). How you know results are good is as important as how you got the results.

This investigation entails using various classes from the library \lstinline{scikitlearn}. To develop the models: \lstinline{GaussianNB} class from the \lstinline{naïve\_bayes} module, the \lstinline{DecisionTreeClassifier} class from the \lstinline{tree} module, the \lstinline{LogisticRegression} class from the \lstinline{linear\_model} module, the \lstinline{KNeighbors} function from the \lstinline{neighbors} module, and \lstinline{LinearDiscriminantAnalysis} class from the \lstinline{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 charactoristics) of each model, \lstinline{sklearn}’s \lstinline{metrics} module was used. Other Python libraries used along the way to help handle the data include \lstinline{matplotlib.pyplot}, \lstinline{pandas}, and \lstinline{numpy}.

\subsection{Working with the Dataset}

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 \lstinline{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 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. The columns and their final values were as follows:

\begin{enumerate}

\item 1. High Blood Pressure (HighBP)

\begin{itemize}

- 0 = no

- 1 = yes

\end{itemize}

2. High Cholesterol (HighChol)

\begin{itemize}

- 0 = no

- 1 = yes

\end{itemize}

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 (HeartDiseaseorAttack)

- 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 = < $10,000

- 2 = < $15,000

- 3 = < $20,000

- 4 = < $25,000

- 5 = < $35,000

- 6 = < $50,000

- 7 = < $75,000

- 8 = > $75,000

\end{enumerate}

Since there were 21 total features in the dataset, to decrease complexity,

After the diabetes dataset processed, it was split into a train group and test group using the \lstinline{model\_selection.test\_train\_split()} function, with 2/3 of the data as the train group and the remainder as the test group.

\subsection{Creating and Comparing the Models}

Next, the models were developed. For the decision tree, logistic regression, and KNN models, similar methods were used as described in previous studies [4]–[6]. Regarding specific parameters, the decision tree model was developed with the Gini impurity criterion, the logistic regression model was developed without any penalty, and the KNN model was developed using k=10 and the Euclidean distance metric. These were determined to be the most effective version of each model in their respective investigation. Additionally, all models were trained using the same training subset of the dataset. For the naïve bayes classifier, the \lstinline{naïve\_bayes.GaussianNB()} class was used and trained on the same training dataset. For the linear discriminant analysis, the \lstinline{linear\_discriminant.LinearDiscriminantAnalysis()} class was used and also trained on the training dataset. For each of these models, the classification report was created using the \lstinline{metrics.classification\_report()} function which displays the accuracy, precision, and recall scores of the model overall and with respect to each classification.

After creating these models, they were applied to the testing subset of the dataset. The classification predictions were then compared to the actual testing data classifications. A confusion matrix was created for each model using the \lstinline{confusion\_matrix()} method from the \lstinline{metrics} module. Using the \lstinline{ConfusionMatrixDisplay} method, also from \lstinline{metrics}, a depiction of the matrix was developed.

To compare 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 model is similar to the methodology explained in a previous report [7]. However, since the iris dataset consists of 3 classifications rather than 2, the data had to be reconfigured to binarize the target using one-hot-encoding. This process uses a “one versus all” computing strategy where a single classification is presented as one class, and all the other classes are presented as another. For this investigation, the class showing the lowest precision values across the models was the versicolor class; this class was used to evaluate the ROC curves. Using the \lstinline{preprocessing.LabelBinarizer()} class, object was fit to the dataset. The object was then able to transform the targets (the true classifications) and predictions (classifications predicted by the model) into the binary classifications where the desired class was represented as 1, and all other classes were represented as 0. Using these transformed arrays, the false positive rate, the true positive rate, and the area under the curve values were calculated for each model using the \lstinline{metrics.roc\_curve()} and \lstinline{metrics.auc()} functions. The false positive and true positives rates from each models were graphed using \lstinline{matplotlib.pyplot} and the AUC values were compared.

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 [8].

# Results

14. Describe computational results; do the answers you got make sense?

15. Describe how a decision would be made given the results; explain cross-validation results, etc.

16. Describe what the results mean in relationship to other data that was presented in the intro.

# Discussion

17. Why are the results what they are, are they what were expected, etc.

18. Why is this work important?

19. What could be done in the future to make things better, or how could this work be extended to cover more data, etc.

Ack & Citations

20. List all acknowledgments and provide citations for all references in the paper. These should be hyper-linked so a reader can click on the citation [n] and go to the reference in the bibliography.