Diabetes Predictive Modeling

ECS 171 Machine Learning
Group 19 Project Report

Group Members:

Justin Nguyen, Daniel Heredia, Ronit Amar Bhatia, Amanda Tu, Katie Sharp

Github Repository:

https://github.com/jstnguyen/DiabetesPredictiveModel

Introduction and background

Diabetes continues to be a life-threatening condition that affects how millions of people process glucose—the cell's primary source of energy. Type 1 Diabetes is an autoimmune condition in which the immune system mistakenly attacks and destroys the insulin-producing beta cells in the pancreas. As a result, people with Type 1 diabetes do not produce insulin and must take costly insulin injections or use an insulin pump to regulate their blood sugar levels. Those with Type 2 Diabetes carry a metabolic disorder characterized by insulin resistance, where the body's cells do not respond effectively to insulin. Over time, the pancreas may also produce less insulin.

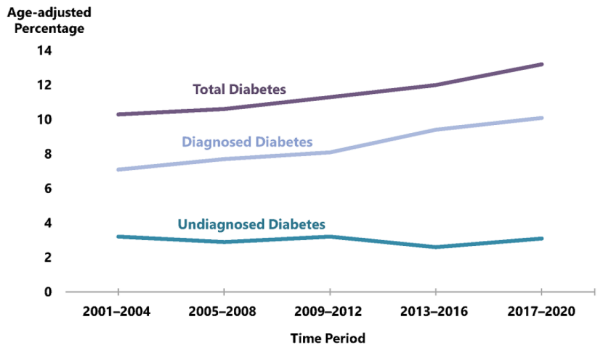

Figure 1: Prevalence of diagnosed, undiagnosed, and total diabetes among U.S. adults

According to the CDC's National Diabetes Statistics Report, Diabetes is "the eighth leading cause of death in the United States, diabetes costs a total estimated $327 billion in medical costs and lost work and wages" [1]. Moreover, 37.3 million people in the United States of America have diabetes and alarmingly 8.5 million people have diabetes but are unaware of their condition due to a lack of diagnosis. Figure 1 graphs the percentage of U. S. adults with undiagnosed diabetes. Diabetes, when left untreated, poses a significant risk to an individual's vision, nervous system, and heart, and even leads to death.

Predictive modeling is prevalent throughout the field of machine learning. Common problems include training models on inaccurate or incomplete errors, overfitting models on the training set, and high model variance. Even with subject matter expertise, human doctors still sometimes fail to diagnose every patient accurately. As a result, predictive models may struggle with achieving accuracy in diagnosis.

If an accurate and reliable model can be built, it can be applied in many practical applications:

- Medical diagnoses
  - 8.5 million people have diabetes but are unaware of their condition due to a lack of diagnosis. A predictive model could help these people receive their diagnoses and begin the treatment they need.
- Financial
  - "Diabetes costs an estimated $327 bil-

lion in medical costs and lost work and wages" [1]. Earlier diagnosis could lead patients to receive the help they need and prevent lost work and wages. Early diagnosis and treatment could also reduce the severity of the condition and the cost of treatment.

Given the inadequacy of current diagnosis methods and diabetes' severe long-term health consequences, we find it worthy and relevant to create a predictive model based on patients' demographic information and medical history.

Literature Review

Machine learning is indeed a disruptive technology benefiting various sectors, including healthcare. For example, machine learning models can analyze X-rays and CT scans, and diagnose illnesses based on a patient's demographic information and medical history. Presented are selected studies revealing how machine learning models can be used to diagnose diabetes in patients.

Researchers Chun-Yang Chou, Ding-Yang Hsu, and Chun-Hung Chou in their paper, "Predicting the Onset of Diabetes with Machine Learning Methods," studied how machine learning models can be used to predict diabetes in Taiwanese citizens [2]. Gathering data from 15,000 women between the ages of 20 and 80 from the Taipei Municipal Medical Center, they compared four models using two-class logistic regression, a two-class neural network, a two-class decision jungle, and a two-class boosted decision tree. They concluded that the two-class decision jungle, which builds from the random forest algorithm using directed acyclic graphs, achieved the best results overall with an AUC of 99.1% [3].

A similar paper, "Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes," by Ravaut et al.achieved substantial progress toward a gradient-boosting decision tree diabetes predictor [4]. Their model processed data from over two million residents from Ontario, Canada, including demographic information, census data, physician claims, laboratory results, prescription medication history, hospital records, ambulatory usage, and more. Their gradient-boosting decision tree model achieved an AUC of 80.26%, demonstrating exceptional ability to distinguish between individuals at high and low risk of diabetes. This research not only stresses the potential of machine learning in diabetes prediction but also highlights the necessity for ongoing research, continual refinement of models, and validation across diverse populations and healthcare systems.

Dataset Description and Exploratory Data Analysis of the Dataset

Our model is trained on the "Diabetes prediction dataset" and is described as a comprehensive dataset for predicting diabetes with medical and demographic data. There are 100,000 rows and 9 columns, including 2 categorical columns and 7 numerical columns. The categorical columns in the dataset are "smoking_history" and "gender." Smoking history included 6 categories indicating the condition of an individual's smoking: "never," "not current," "former," "current," "ever," and "No Info." "Former" and "not current" differ in how "former" refers to individuals that have refrained from smoking for a longer period of time than those that are "not current" smokers. We have decided to remove the "smoking_history" column from training our model since more than 33,000 observations report "No Info." "Gender" is limited to female and male, with 18 observations reporting "Other. We have decided to omit these 18 observations due to their limited availability in data. The numerical columns in the dataset are "age," "hypertension," "heart_disease," "bmi," "HbA1c_level," "blood_glucose_level," and "diabetes." "Hypertension" and "heart_disease" refer to a medical condition in which the blood pressure in the arteries is persistently elevated and any disease pertaining to the heart including diseased vessels, heart structural problems, blood clots, and more. A value of 1 indicates the presence of the condition and a value of 0 means the absence of the condition. "Age" and

"bmi" refer to a person's biological age and body mass index. "HbA1c_level" is a measure of a person's average blood sugar level over the past 2 to 3 months. "Blood_glucose_level" refers to the amount of glucose in the bloodstream at a given time. Finally, this dataset is labeled with the "diabetes" column with 1 indicating presence of diabetes and 0 indicating the absence of diabetes. Categorical data was encoded using one-hot encoding and numerical data was scaled using min-max scaling.

An important part of the exploratory data analysis involved learning about the trends in the data. We began by eliminating all duplicates in the dataset. Due to the dataset's large size, the goal was to exclude unnecessary rows. The examination of each feature followed, with an analysis of the spread of values and addressing anything that caught our attention. Following this, a correlation matrix was generated. No strong correlations stood out, except for the relationship between diabetes and HbA1c levels, which is logical given that HbA1c measures the amount of sugar in red blood cells. A pair plot was also created, revealing a distinct separation concerning individuals with and without diabetes.

Proposed Methodology

Diabetes prediction is a binary classification problem. Based on our Literature Review, we concluded that logistic regression, Naïve Bayes, and random forest models are best suited for binary classification problems. Our research revealed that logistic regression is a powerful predictor with a balance between high accuracy and simplicity. Additionally, we selected Naïve Bayes for its speed and effectiveness in handling categorical data, and random forest because it is a high-accuracy model that performs well with minimal tuning.

We will start our data preprocessing by removing the "smoking_history" column and rows with "Other" for "gender." We justify this removal because of their lack of data. We then perform one-hot encoding on categorical columns "gender," "hypertension," "heart_disease," and "diabetes." We conclude preprocessing by normalizing the rest of the data using min-max scaling.

Our models are evaluated using k-fold cross-validation to ensure that our findings are robust and applicable beyond our dataset. Accuracy, precision, recall, and the F1 score will be our evaluation metrics of choice. Additionally, Receiver Operating Characteristic (ROC) curves will be used to calculate the Area Under Curve (AUC) scores to see how well our models distinguish between patients with and without diabetes across various thresholds.

To create our model, we will be using Python's Pandas and scikit-learn libraries for their powerful data processing and machine learning capabilities. Python's pickle package serializes our logistic regression model which allows it to be called by our HTML-based webpage. Our webpage prompts users to input medical data and returns a prediction of whether or not diabetes may be present. Once we have run our models and gathered the data, we will analyze the outcomes, measure them against existing research, and share our discoveries—aiming to contribute valuable insights to the ongoing conversation around diabetes prediction.

Experimental Results

Logistic Regression:

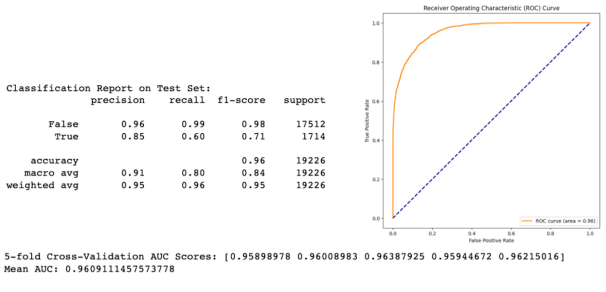

Figure 2: Testing Results of Logistic Regression

For our first diabetes prediction model, we utilized a logistic regression model, which is typically used for binary classification tasks pre-

dicting the probability of an instance belonging to one of two classes. Our logistic regression model demonstrated a robust discriminative ability with an accuracy of 96.0%.

Delving into the negative (False) class, we observe high scores with 96.0% precision and 99.0% recall. These findings reveal that our logistic regression model is highly effective in correctly identifying instances of the negative class (when a person does not have diabetes) with a precision of 96.0%. This minimizes false positives which would have led to costly overdiagnosis and overmedication. Furthermore, the high recall of 99.0% for the False class underscores the model's reliability in capturing the majority of actual negative instances. However, the positive (True) class' precision and recall scores at 85.0% and 60.0%, respectively, suggest some room for improvement enhancing the model's ability to correctly identify positive instances. Improving how our model predicts the positive class is especially important in reducing false negatives which could become fatal as diabetes is left undiagnosed and untreated.

Finally, we assessed the model's generalizability through a 5-fold cross-validation process, yielding a Mean AUC score of 96.1%. This consistency highlights the model's reliable performance, implying the model is robust and not overly dependent on random subsets of data.

We will eventually see that logistic regression was chosen as our best model based on accuracy and computational demand. Logistic regression achieves a Mean AUC of 96.1% which is only 0.2% less than random forest's AUC of 96.3%. We are comfortable with this marginal performance tradeoff because our logistic regression model trains significantly faster than our random forest model—making it a reliable and practical model.

Naïve Bayes:

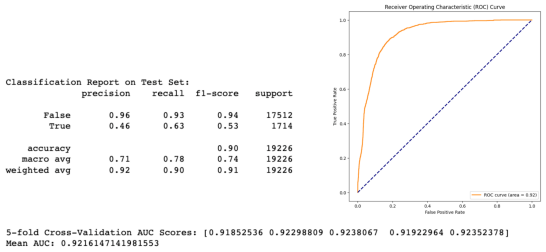

Figure 3: Testing Results of Naïve Bayes

For our second diabetes prediction model, we implemented the Naïve Bayes classification algorithm. Naïve Bayes assumes that features used to describe an instance are conditionally independent, given the class label. This assumption works well in practice and makes the model significantly more computationally efficient. It then calculates the probability of an instance belonging to a class given its features. In this case, it is the probability of a patient having diabetes or not, given their medical data. Our Naïve Bayes model achieves modest discriminative ability with an accuracy of 90.0%.

We note that the model achieves high precision and high recall for the negative class with 96.0% and 93.0% respectively, indicating exceptional ability to identify true negatives. However, we also note that the model struggles with the positive class significantly more than our logistic regression model. Our Naïve Bayes model achieves a precision of 46.0% and recall of 63.0%, suggesting that the model will struggle with correctly identifying positive instances. Therefore, this model is unreliable in capturing the positive class and could potentially produce fatal false negatives.

Finally, we assessed the model's generalizability through a 5-fold cross-validation process, yielding a Mean AUC score of 92.2%. This consistency highlights the model's reliable performance, implying the model is robust and not overly dependent on random subsets of data.

Reflecting on our approach, the Naïve Bayes

classifier was selected for its simplicity and efficiency, particularly suitable for our dataset's high dimensionality. While the model exhibits a strong foundation in identifying negative non-diabetic instances, the precision and recall for positive diabetic predictions suggests potential areas for further refinement.

Random Forest:

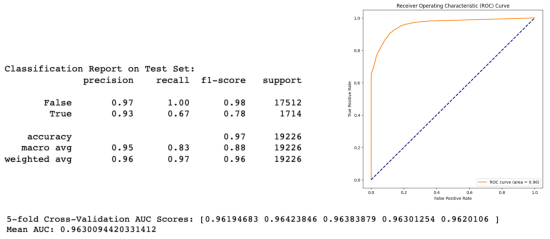

Figure 4: Testing Results of Random Forest

Our third model implements the random forest algorithm. Random forest creates multiple decision trees through bootstrap sampling where each tree will vote for a class. Predictions from all trees are averaged for a final input. Our random forest classifier is initialized with 200 trees and achieves robust discriminative ability with an accuracy of 97.0%.

The model performs well for the negative class achieving both high precision and high recall with 97.0% and 100.0% respectively. This model has exceptional ability to reliably identify true negatives, and capture the negative class. We also note that the model has impressive precision for the positive class with 93.0%, revealing that it will accurately create positive predictions. Its recall of 67.0%, however, suggests that the model does not effectively capture a significant portion of positive instances. This could be a fatal mistake as diabetes goes undiagnosed.

Finally, the model achieved consistent performance with a Mean AUC of 96.3% across 5 folds of k-folds cross-validation, confirming the model's resilience across variations in the dataset.

Although the random forest model achieved impressive results, we note that this random forest is a computationally intensive method and impractical to use for our purposes.

Conclusion and Discussion

Utilizing logistic regression, our model achieved an impressive AUC of 96.1%. This performance stands favorably in comparison to Naïve Bayes, with an AUC of 92.2%, and random forest which boasts an AUC of 96.3%. Despite a marginal 0.2% AUC difference between random forest and logistic regression, we have chosen logistic regression as our primary approach. This decision is justified by its notably high AUC and its considerable speed advantage over the random forest in execution.

Turning our attention to the dataset, room for improvement is evident. Currently, our model is exclusively trained on male and female observations due to a lack of nonbinary data. To enhance the real-world applicability of our model, we recommend increasing nonbinary observations. Also recognizing the established association between smoking and the risk of developing diabetes, we suggest the collection of additional smoking history data. Currently, our dataset's labels do not specify between Type 1 Diabetes or Type 2 Diabetes. We propose collecting data that specifies between Type 1 Diabetes and Type 2 Diabetes in order to develop a model that makes more specific predictions in regards to diabetes presence and its type. Elevating the quality of our dataset in these dimensions will increase the generalizability and accuracy of our model, consequently amplifying the reliability of its predictions and the potential for life-saving impact.

Meeting Date	Milestones	Findings
10/11/2023	• Ideation • Diabetes literature review	We decided to develop a model in order to predict whether or not a patient will have diabetes, a condition that affects insulin production and response to insulin.
10/18/2023	• Completion of One-Pager • Predictive models literature review	We reviewed two papers. In their article "Predicting the Onset of Diabetes with Machine Learning Methods" (Chou et al., 2023), researchers Chun-Yang Chou, Ding-Yang Hsu, and Chun-Hung Chou created a neural network. Researchers Ravaut et al. also created a gradient-boosting decision tree model in their paper, "Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes" (2021).
11/01/2023	• Exploratory data analysis • Removal of outliers • Selection of logistic regression as the proposed methodology	We justified the removal of the "Other" category in "gender" because there were only 17 observations out of 100,000 total observations. We also excluded the "smoking_history" column from our data because more than 33,000 observations did not include smoking history. We performed One-Hot Encoding on categorical columns like "hypertension," "'heart_disease," "gender," and "diabetes." We scaled all data with Min-Max Scaling. We found no outliers to remove.
11/15/2023	• Completion of predictive models • Evaluation of model performance • Confirmed selection of logistic regression	We completed and tested 3 predictive models: logistic regression, Naïve Bayes, and random forest. We used K-Folds Cross Validation and Area Under Curve (AUC) in order to validate and compare these models to select the best model option. We found that logistic regression had an AUC of 96.09%, Naïve Bayes had an AUC of 92.16%, and random forest had an AUC of 96.31%. We chose logistic regression on the basis of accuracy and computational demand. Logistic regression has one of the highest AUC which is comparable to random forest's AUC, but finishes its execution significantly faster.
11/17/2023	• Began to create HTML-based frontend	We proposed a website that receives medical data input, calls the logistic regression model, and outputs a prediction based on the inputted data.
11/22/2023	• Added finishing touches to the website and report	We added aesthetic design to the website and debugged an input scaling issue. We completed our project report with its conclusion and discussion.
11/29/2023	• Final submission	We prepared a PowerPoint, and video demo, and updated the README.md.

References

[1] Centers for Disease Control and Prevention (2023). National Diabetes Statistics Report. National Diabetes Statistics Report website. https://www.cdc.gov/diabetes/data/statistics-report/index.html. Accessed 11/15/2023.

[2] Chou, Y., Hsu, Y., & Chou, H. (2023). Predicting the Onset of Diabetes with Machine Learning Methods. Journal of Personalized Medicine, 13(3). https://doi.org/10.3390/jpm13030406

[3] Shotton, J., et al. (2013). Decision Jungles: Compact and Rich Models for Classification. Microsoft Research. https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/DecisionJunglesNIPS2013.pdf

[4] Ravaut, M., Harish, V., Sadeghi, H., et al. (2021). Development and Validation of a Machine Learning Model Using Administrative Health Data to Predict Onset of Type 2 Diabetes. JAMA Network Open, 4(5), Article e2111315. https://doi.org/10.1001/jamanetworkopen.2021.11315