# Diabetes Classification Analysis

Joshua Vilela, Tristan Chen, Matthew Xu

Joshua Vilela: Data background and Discussion
Tristan Chen: Result Analysis and Interpretation
Matthew Xu: Coding and Methodology

## 0. Abstract

For our final project, we have decided to study the effects of human characteristics and conditions to predict and classify the presence of diabetes. Diabetes is one of the largest epidemics in human history, and it continues to grow. If left untreated, this disease can cause permanent damage to the kidneys, eyes, and heart. In order to control this epidemic, many researchers have been working on models for early detection of diabetes. Our objective is to create a model that details which conditions may lead to a higher risk of diabetes. We decided to model the data using logistic regression, which showed that polyuria and polydipsia are strong indicators of increased diabetes risk. The model is also fairly accurate with a precision score of 94% and recall score of 93% for correct positive and negative classifications, both of which are fairly high.

---

## 1. Background: Diabetes

Diabetes has become one of the most common life threatening diseases affecting an estimated 461 million people across the globe. In 2019, diabetes was the cause of an estimated 1.5 million deaths. This disease also has a long asymptomatic phase which makes it very difficult to detect. According to the CDC, 34.2 million Americans suffer from diabetes, with 1 in 5 unaware of their condition. The diabetes epidemic is also a large economic burden, as the annual cost associated with the condition is estimated to be more than $300 billion a year.


Early detection of diabetes is crucial for reducing the risk of life-threatening complications, including heart disease, stroke, and kidney failure. In order to help detect diabetes at an earlier stage, researchers have begun to use machine learning methods to create diabetes risk prediction models.

As more data is gathered from diabetic or prediabetic patients, the models are able to evaluate the risk with more precision. These models analyze pre-existing conditions and symptoms that are associated with an increase in diabetes risk, such as obesity, muscle stiffness, weakness, or excess urine.

The data are publicly available on the UCI data repository:

> Islam, MM Faniqul, et al. 'Likelihood prediction of diabetes at early stage using data mining techniques.' Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.

---

# 2. Data Description

This dataset contains the sign and symptom data of newly diabetic or would-be diabetic patients along with patient demographics, collected from the Sylhet Diabetes Hospital in Sylhet, Bangladesh.

## Sample and Measurement Information

This dataset contains data on 17 attributes of 520 patients of the Sylhet Diabetes Hospital in Sylhet, Bangladesh. These attributes include possible symptoms or conditions that the patient may have. This data was collected using direct questionnaires.

## Data Structure

In this study, the observational units are patients at the Sylhet Diabetes Hospital and the variables are the age, pre-existing conditions or symptoms, and a class stating whether the patient is positive or negative for diabetes. With the exception of age, every variable is a two level factor. For example, Sex is a 2 level factor of male or female, and each variable describing a condition is a 2 level factor of yes or no. The population of this dataset is all patients in hospitals, while the sampling frame includes all patients in hospitals that have or are at risk of diabetes. The sample includes patients in the Sylhet Diabetes Hospital who have taken the questionnaire. The scope of inference is small, due to the sample being from only one hospital. As a result, the dataset may not be a great representation of the population.

**Table 1**: Variable Descriptions

| Variable | Description | Units |
|---|---|---|
| Age | Age value in range between 16-90 | Years (yr) |
| Sex | Gender category of patient | Male/Female |
| Polyuria | Whether patient has a history of *Polyuria* | Yes/No |
| Polydipsia | Whether patient has a history of *Polydipsia* | Yes/No |
| Sudden weight loss | Whether patient has the symptom of sudden weight loss | Yes/No |
| Weakness | Whether patient has a history of weakness | Yes/No |
| Polyphagia | Whether patient has a history of *Polyphagia* | Yes/No |
| Genital thrush | Whether patient has a history of Genital thrush | Yes/No |
| Visual blurring | Whether patient has the symptom of visual blurring | Yes/No |
| Itching | Whether patient has the symptom of itching | Yes/No |
| Irritability | Whether patient has the symptom of irritability | Yes/No |
| Delayed Healing | Whether patient has the symptom of delayed healing | Yes/No |
| Partial Paresis | Whether patient has the symptom of partial paresis | Yes/No |
| Muscle sti ness | Whether patient has the symptom of muscle stiffness | Yes/No |
| Alopecia | Whether patient has a history of *Alopecia* | Yes/No |
| Obesity | Whether patient has a history of Obesity | Yes/No |
| Class | Result of tested patient for Sylhet Diabetes | Positive/Negative |

During data processing, we converted the factors into numbers, with 0 representing No, Male, and Negative, and 1 representing Yes, Female and Positive.

**Table 2**: Example rows of data after processing

| | Age | Gender | Polyuria | Polydipsia | sudden weight loss | weakness | Polyphagia | Genital thrush | visual blurring | Itching | Irritability | delayed healing | partial paresis | muscle stiffness | Alopecia | Obesity | class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 40 | Male | No | Yes | No | Yes | No | No | No | Yes | No | Yes | No | Yes | Yes | Yes | Positive |
| 1 | 58 | Male | No | No | No | Yes | No | No | Yes | No | No | No | Yes | No | Yes | No | Positive |
| 2 | 41 | Male | Yes | No | No | Yes | Yes | No | No | Yes | No | Yes | No | Yes | Yes | No | Positive |
| 3 | 45 | Male | No | No | Yes | Yes | Yes | Yes | No | Yes | No | Yes | No | No | No | No | Positive |
| 4 | 60 | Male | Yes | Yes | Yes | Yes | Yes | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Positive |

During initial exploratory analysis, the age distribution of the patients was found to be generally unimodal, with many of the patients being between the ages of 35 and 50. Analysis also showed that there was a much smaller number of female patients that tested negative for diabetes. However, this may simply be a consequence of the fact that this is not a large dataset, and that the data is only gathered from a single hospital.

# Methods

In order for the data to be interpreted and analysed using models for classification, variables such as the labels, gender, and others must be converted to binary for assessment. The first plot visualized is the correlation heatmap to visualize the correlations between individual variables and look at their strengths and spot potential high impact variables. Then, the data was split into training and test sets in a 70 to 30 percent ratio to ensure that both sets have enough data for precise classifications.

Logistic regression was performed on the processed dataset to identify which impacts of each variable on diabetes risk. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. The logistic function, also known as the sigmoid function, which is a function in base e that maps values to between 0 and 1, but never exactly at those limits. Logistic regression is a linear method, but the predictions are transformed using the logistic function.

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

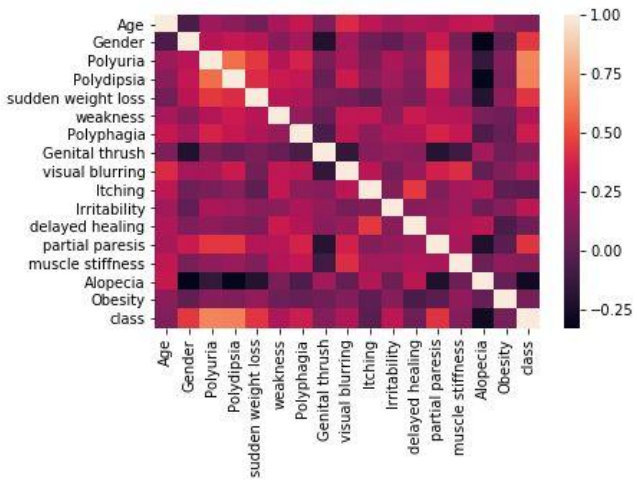
The model was fitted using training data and test data was used to predict the labels for metrics. Variables with higher coefficient estimates are associated with a larger impact on risk. Finally, an ROC curve was created to visualize the effectiveness of logistic regression in modeling this dataset.

The main metric used here is precision-recall, as we are concerned with increasing true positives while decreasing false negatives. This will improve our precision, the reproducibility of the model, and recall, the correct classifications of the model's predictions.

---

# Results

## Variable Correlation




From the heatmap, many strong correlations between two variables lie between two medical conditions, specifically similar ones. As shown above, the lighter the block color, the stronger the correlation is. For example, pairs of illnesses such as Polyuria and Polydipsia have high correlation.

## Logistic Regression

To perform logistic regression, the data was first split into training and test sets for the predictor and response variables. The training set was then used to fit the model. The response variable was class, which described whether a patient tested positive or negative for diabetes. The predictor variables include age, gender, and medical conditions. The results of the analysis are shown in the table below.

**Table 3**: Coefficient estimates for each variable estimated through logistic regression.

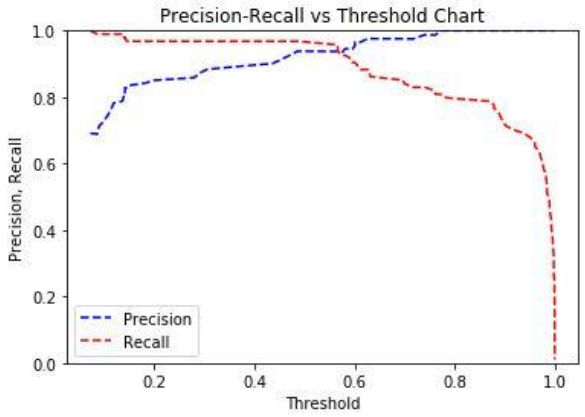
| | 0 | 1 |
|---|---|---|
| 0 | Age | -0.029355 |
| 1 | Gender | 2.600933 |
| 2 | Polyuria | 2.589485 |
| 3 | Polydipsia | 2.752740 |
| 4 | sudden weight loss | 0.572163 |
| 5 | weakness | 0.350271 |
| 6 | Polyphagia | 0.562261 |
| 7 | Genital thrush | 1.091966 |
| 8 | visual blurring | 0.121674 |
| 9 | Itching | -1.066799 |
| 10 | Irritability | 1.526707 |
| 11 | delayed healing | -0.381376 |
| 12 | partial paresis | 0.867519 |
| 13 | muscle stiffness | -0.017095 |
| 14 | Alopecia | -0.193980 |
| 15 | Obesity | 0.026801 |
| 16 | class | NaN |

The coefficient estimates shown above represent the impact that the presence of each condition has on the overall risk for diabetes. Positive coefficients represent a positive impact, while negative coefficients represent a negative impact. Each number represents a multiplicative increase in the log odds of having diabetes when the condition is present. Polydipsia and Polyuria have the largest positive coefficients, so it can be assumed that these conditions are strong indicators of diabetes risk. Other coefficients that are associated with a higher diabetes risk include sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, irritability, partial paresis, and obesity.

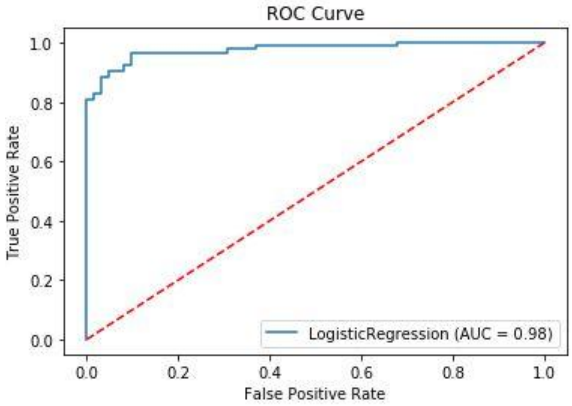
Table 3 also shows the coefficient estimates for demographic factors, including age and gender. The estimates for age and gender are -0.029 and 2.601 respectively, indicating that age is negatively associated with diabetes risk and being female is positively associated with diabetes risk. This does not follow the results from other studies, which suggest males and elderly have a higher risk of the disease. This disparity may be a result of the small size of the sample.

## Accuracy of the Model

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.90 | 0.92 | 62 |
| 1 | 0.94 | 0.96 | 0.95 | 94 |
| accuracy | | | 0.94 | 156 |
| macro avg | 0.94 | 0.93 | 0.93 | 156 |
| weighted avg | 0.94 | 0.94 | 0.94 | 156 |




The model has a decently high precision-recall score, as both scores are about 94% for either classification of negative or positive presence of diabetes. The precision-recall plot shows the impact of the probability threshold used on the model. As the threshold increases, precision continues to increase while recall decreases. The objective of this plot is to show the optimal prediction threshold, which in this case is about 0.6 probability for classification.




To determine the accuracy of the model, we first set the test set against our fitted model. This gave us a mean accuracy score of 0.94. We then created an ROC curve, which plots the TPR and FPR of the model. As shown in figure 3 below, the AUC (Area Under Curve) is 0.98, indicating that the model performs very well.

# Discussion

### Which medical conditions and symptoms have a larger impact on diabetes risk?

The coefficient estimates shown above represent the impact that the presence of each condition has on the overall risk for diabetes. Positive coefficients represent a positive impact, while negative coefficients represent a negative impact. Each number represents a multiplicative increase in the log odds of having diabetes when the condition is present. Polydipsia and Polyuria have the largest positive coefficients, so it can be assumed that these conditions are strong indicators of diabetes risk. Other coefficients that are associated with a higher diabetes risk include sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, irritability, partial paresis, and obesity. Each coefficient can be interpreted similarly to the following example; the presence of polyuria in a patient is associated with a multiplicative increase in the odds of having diabetes.

### Do demographic factors such as age and gender impact having a higher risk of diabetes?

Table 3 also shows the coefficient estimates for demographic factors, including age and gender. The estimates indicate that age is negatively associated with diabetes risk and being female is positively associated with diabetes risk. This does not follow the results from other studies, which suggest males and elderly have a higher risk of the disease. This disparity may be a result of the small size of the sample.

Other things that we can explore is delving into ethical bias and see how much of an impact it would have in the risk of diabetes alongside age and gender. The data stems from one hospital at a particular location, possibly having ethical bias as the population has high density for its sample. There are also many other factors that are hidden that we could try and look into as these hidden factors could play a significant role in finding the risk of diabetes. There are also many physical factors that could be a big part such as the blood pressure of the patients or their respective weight which could further understand how we can determine the risk in the patients.