RIT | Rochester Institute of Technology

INVESTIGATING THE RISK FACTORS FOR DIABETES IN PIMA INDIANS

APPLIED STATISTICS

PROJECT REPORT

-DEVIKA PILLAI
MS IN DATA SCIENCE

1. INTRODUCTION

Diabetes is a chronic metabolic illness that affects millions of people throughout the world. Understanding the factors that contribute to the genesis of this disease is critical for prevention and treatment. The Pima Indians Diabetes Dataset, which is originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains information on 768 women of Pima Indian heritage aged 21 years and above. The objective of this dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. The dataset contains numerous characteristics such as age, BMI, blood pressure, insulin level, and so on, as well as a binary target variable indicating whether the patient has diabetes. The dataset is suitable for binary classification problems and is frequently used to train and test machine learning models for predicting the likelihood of diabetes in patients based on diagnostic metrics.

In this project, I aimed to analyze the Pima Indians Diabetes Dataset to identify the risk factors associated with diabetes using statistical analyses. To achieve this, I formulated four research questions that I aimed to address through my analysis. Overall, this project aims to contribute to the understanding of risk factors associated with diabetes and to provide insights that could aid in the prevention and treatment of this chronic metabolic disorder. Furthermore, the project's findings have the potential to impact public health policies and programs focused at reducing the incidence and prevalence of diabetes. By recognizing the risk factors for diabetes, healthcare providers and policymakers can establish targeted interventions and educational programs to help individuals lower their chance of developing the condition. Furthermore, the findings of this investigation could be used to influence the development of innovative diagnostic and treatment approaches that could improve patient outcomes and lower the overall burden of diabetes on individuals and society.

2. METHODS

The Pima Indians Diabetes Dataset is a publicly available dataset that is widely used in machine learning and data analysis projects. It only consists of numerical data and the dataset includes 9 variables, which are as follows:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure: Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skinfold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/(height in m)^2)
- DiabetesPedigreeFunction: Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
- Age: Age in years
- Outcome: The binary target variable indicating whether the patient has diabetes (0 = No diabetes, 1 = Diabetes)

The dataset has no missing values and is relatively balanced, with 268 patients without diabetes and 500 patients with diabetes. Except for the variables SkinThickness and Insulin, which are positively skewed, the data is mainly normally distributed.

2.1. Are the average glucose levels in the population significantly different from the recommended levels (100 mg/dL or below)?

To address this question, I will use a one-sample t-test, where the null hypothesis is that the mean glucose level of the population is equal to 100 mg/dL. If the p-value comes out to be less than the significance level, then I will reject the null hypothesis. This test can provide insights into the overall health of the population with respect to their glucose levels and whether the population needs to make changes to their diet and lifestyle to manage their glucose levels.

2.2. Is there a significant difference in glucose levels between women with and without diabetes?

For this question, I will use a two-sample t-test, where the null hypothesis is that the mean glucose levels between women with and without diabetes. If the p-value comes out to be less than the significance level, then I will conclude that there is a significant difference in glucose levels between the two groups. This test can provide insights into the relationship between diabetes and glucose levels in women and help identify whether glucose levels are a useful biomarker for identifying diabetes.

2.3. Can we predict the likelihood of developing diabetes based on age, BMI, insulin level, and glucose level?

I will use multiple linear regression analysis for this question. The dependent variable would be the likelihood of developing diabetes, and the indecent variables would be Age, BMI, Insulin level, and Glucose level. The regression model will help identify which of these variables have a significant impact on the likelihood of developing diabetes and how they interact with each other.

2.4. Is there a significant difference in BMI among women with diabetes who have had one, two, or three or more pregnancies?

For this question, I will use one-way ANOVA, where the null hypothesis is that the mean BMI of women with diabetes is the same for those who have had one, two, or three or more pregnancies. If the p-value is less than the significance level, then I will conclude that there is a significant difference in BMI among the different groups of women.

3. RESULTS

3.1. Are the average glucose levels in the population significantly different from the recommended levels (100 mg/dL or below)?

H_0: $\mu = 100$
H_1: $\mu \neq 100$
$\alpha = 0.05$
$n = 768$
$\nu = n - 1 = 767$
$\bar{X} = 120.8945$
$t_0 = 18.111$

p-value < 2.2e-16
The 95% confidence interval lies between 118.6297 and 123.1593.
Decision: Since the p-value is less than α, we have enough evidence to reject H_0, hence our decision is to reject H_0.
Conclusion: We can conclude at 5% significance level, that the average glucose levels in the population is significantly different from the recommended level of 100 mg/dL or below.

3.2. Is there a significant difference in glucose levels between women with and without diabetes?

H_0: $\mu_{without} - \mu_{with} = 0 \Rightarrow \mu_{without} = \mu_{with}$
H_1: $\mu_{without} - \mu_{with} \neq 0 \Rightarrow \mu_{without} \neq \mu_{with}$
$\alpha = 0.05$
n = 768
$v = 461.33$
$\bar{X}_1 = 109.9800$
$\bar{X}_2 = 141.2575$
$s_1 = 26.1412$
$s_2 = 31.93962$
$t_0 = -13.752$
p-value < 0.00000000000000022
The 95% confidence interval lies between -35.74707 and -26.80786.
Decision: Since the p-value is less than α, we have enough evidence to reject H_0, hence our decision is to reject H_0.
Conclusion: We can conclude at 5% significance level, that the average glucose levels for women without diabetes is significantly different from the average glucose levels for women with diabetes.

3.3. Can we predict the likelihood of developing diabetes based on age, BMI, insulin level, and glucose level?

The below image shows the multiple linear regression model summary. In it, the intercept, Glucose, BMI, and Age are significant variables as their p – values are lower than α. Intercept and Insulin have a negative effect on the diabetes outcome, but the other 3 variables have a positive effect. This means as the insulin increases, the diabetes value decreases but as BMI, Glucose or age increases, the likeliness of having diabetes also increases. Yes, we can predict the likelihood of developing diabetes based on age, BMI, insulin level, and glucose level. The AUC (Area under the curve) of the model is 0.8248, which means the model performs well and predicts the right output almost always. The R^2 value is 0.2758 which means that only about 27.58% of the variation in outcome can be explained by the model and the MSE is 0.4079. The model is also significant as the p – value is smaller than α.
The fitted regression equation is:
Outcome = -0.9453469 + 0.0061189*Glucose + 0.0123781*BMI -0.0001187*Insulin + 0.0050535*Age

```
Call:
lm(formula = Outcome ~ Glucose + BMI + Insulin + Age, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95045 -0.29451 -0.09727  0.33709  1.25541

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -0.9453469  0.0926808 -10.200 < 0.0000000000000002 ***
Glucose      0.0061189  0.0005845  10.469 < 0.0000000000000002 ***
BMI          0.0123781  0.0022003   5.626          0.0000000282 ***
Insulin     -0.0001187  0.0001524  -0.779             0.436427
Age          0.0050535  0.0014459   3.495             0.000508 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4079 on 610 degrees of freedom
Multiple R-squared:  0.2758,    Adjusted R-squared:  0.271
F-statistic: 58.07 on 4 and 610 DF,  p-value: < 0.00000000000000022
```

Fig.1: Multiple linear regression model

```
Data: pred in 101 controls (test$Outcome 0) < 52 cases (test$Outcome 1).
Area under the curve: 0.8248
```

Fig.2: AUC results

The normal probability plot shows that they are the residuals deviate from the straight line hence it is not normally distributed, and it is a light tailed distribution. The residuals vs fitted values plot shows that the values lie in 2 diagonal lines which mean that the relationship between the independent variables and the dependent variable may not be linear. This means that the model may not capture the true relationship between the variables, which can lead to biased or incorrect predictions.

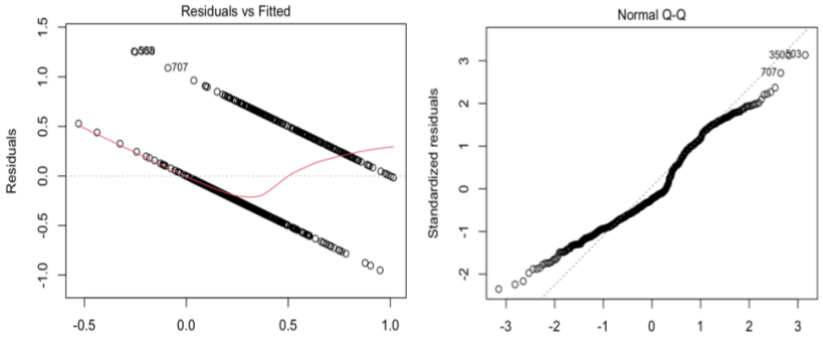

Fig.3(a) Residuals vs fitted plot; (b)Normal probability plot of residuals

3.4. Is there a significant difference in BMI among women with diabetes who have had one, two, or three or more pregnancies?

The below image shows the results for performing ANOVA to see if there is a significant difference in BMI among women with diabetes who have had one, two, or three or more pregnancies.

```
Analysis of Variance Table

Response: BMI
           Df  Sum Sq Mean Sq F value  Pr(>F)
Preg        2   373.7 186.859  4.0457 0.01877 *
Residuals 227 10484.5  46.187
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig.3: ANOVA results

H_0: $\mu_{one} - \mu_{two} = \mu_{threeormore}$
H_1: At least one mean is different from the others.
There are 3 levels(treatments)
a = 3
There are 29 replicates for the 1st level, 19 replicates for the 2nd level and 182 replicates for the 3rd level.
The p – value is less than α, therefore we have enough evidence to reject H_0. Hence, we can claim that there is a significant difference in BMI among women with diabetes who have had one, two, or three or more pregnancies.

4. CONCLUSION & RECCOMMENDATIONS

In conclusion, I aimed to analyze the Pima Indians Diabetes Dataset in my project to study the risk factors associated with diabetes using statistical analyses using four research questions. The results of the analyses provided insights into the relationships between various factors and diabetes risk and could help build public health policies and interventions to reduce the incidence and prevalence of diabetes. Based on the first and second statistical analysis conducted, we can conclude that the average glucose levels in the population are significantly different from the recommended levels of 100 mg/dL and that there is a significant difference in glucose levels between women with and without diabetes. Based on the multiple linear regression analysis, the model shows that Glucose, BMI, and Age have a positive effect on the diabetes outcome, whereas Insulin has a negative effect. The model has an AUC of 0.8248, indicating good predictive performance and, the model is significant. The model only explains about 27.58% of the variation in the diabetes outcome, indicating that there are other factors that may contribute to the development of diabetes that are not captured in the model. The model may not be able to capture the true relationship between the variables, which can lead to biased or incorrect predictions. In such cases, one possible solution is to consider using a different regression model that better fits the data or to transform the data to achieve a linear relationship.

Based on the ANOVA results, we can conclude that there is a significant difference in BMI among women with diabetes who have had one, two, or three or more pregnancies. The dataset suggests that in the real-world, the population could have elevated glucose levels, which could increase the risk of developing diabetes and other health issues. We could conduct further research to explore the relationship between glucose levels and diabetes in different populations and investigate effective interventions for managing diabetes. The multiple linear regression analysis can be used to identify individuals at high risk of diabetes based on age, BMI, insulin level, and glucose level. Recommendations could include investigating which group has significantly different BMIs and considering the impact of pregnancy history on diabetes management in women. Healthcare providers should discuss pregnancy history as part of routine diabetes management and consider its potential impact on treatment decisions.