Predicting Diabetes Risk with Multiple Linear Regression Analysis

COHNDSE23.3F-019

OSHADHA KAVEEN SAMARASINGHE

Contents

[Introduction 3](#_Toc166360714)

[Data Exploration and Visualization 4](#_Toc166360715)

[Data Preprocessing 6](#_Toc166360716)

[Linear Regression Model 7](#_Toc166360717)

[Model Evaluation 9](#_Toc166360718)

[Prediction 10](#_Toc166360719)

[Conclusion 11](#_Toc166360720)

[Code Implementation 12](#_Toc166360721)

# Introduction

The purpose of this analysis is to predict the risk of diabetes using multiple linear regression in statistics. Multiple linear regression model is used when we want a prediction and conclusion from a dataset. Multiple linear regression is used to model the relationship between dependent and independent variables. According to this scenario the target(dependent) variable is the risk of diabetes. Features(independent) variables are containing the health-related factors such as smoking, sex, serum creatinine, serum sodium, high blood pressure, platelets, ejection fraction, age, anaemia, diabetes, and creatinine phosphokinase.

# Data Exploration and Visualization

First, we want to load the data set correctly. After inserting we can view the dataset. There are 11 variables and 299 observations. These includes diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, diabetes, creatinine phosphokinase, anaemia, age.

Data Visualization is used to identify the relationship between target variable and features. For the data visualization we can use scatter plots, density plots, box plots, gg plots. In this case I used all of these to show the relationship between ejection fraction, serum sodium, platelets, serum creatinine and creatinine phosphokinase.

# Data Preprocessing

In data preprocessing we must check whether dataset has any missing values. According to our dataset there were no missing values. Because of that there is no need for data preprocessing.

We must split our dataset into 80:20 ratio before building the predictive model. Use caTools library for the split part. From this we can get an idea about how our model’s performance on unseen data. After splitting part, we can separate 80% dataset as Train dataset and 20% dataset as Test dataset.

# Linear Regression Model

In this part we implement a linear regression model. For this implementation we use our 80% train dataset. Training data helps us to train our model. When comes to Linear regression model there are some points that need to be satisfied.

1) Y must be numerical

2) X can be numerical or cat-ordinal

3) Y must depend on X

Our model mainly aimed to predict ejection fraction based on platelets, creatinine phosphokinase, and serum sodium levels.

According to our model Coefficients,

Intercept: -15.99

Platelets: 6.201

Creatinine phosphokinase: 3.804

Serum Sodium: 3.789

Below part shows the equation for multiple linear regression that we use based on this scenario.

# Model Evaluation

Model performance is evaluated by using Mean Square Error (MSE), Root Mean Squared Error (RMSE), and R Squired.

Mean Squared Error: 159.4595

This value represents the average difference between actual and predicted values. Predictive performance is better if MSE value contains lower value. In our case MSE value contains high value. So, our model predictions may have relatively high errors.

Root Mean Squared Error: 12.62773

This value represents the square root of the MSE and provide an estimate of the standard deviation of prediction error. Prediction errors are low if RMSE contain lower value. According to our model it contains high value, so our model has relatively high errors.

R-Squared: 0.02467521

This value represents the proportion of variance in the dependent variable that explained by independent variable. In good model R squared value greater than 50%. According to our model it is about 2.46% So without doubt we can identify our model has relatively high errors.

# Prediction

When we come for prediction part predictions were made by using test data using trained model. We can identify the difference between actual and predicted values.

Below proofs shows when print and summary, our data comparison. Additionally, we visualize comparison between our actual and predicted values using scatter plot and histogram.

# Conclusion

In conclusion, this multiple linear regression model has low performance on predicting the risk of diabetes. When considering MSE, RMSE, R-Squared value we can clearly see our model prediction has a relatively high error. Finally, we can consider our multiple linear regression model is not suitable for predicting the risk of diabetes.

# Code Implementation