NGHIÊN CỨU XÂY DỰNG MÔ HÌNH HỒI QUY LOGISTIC BẰNG PHẦN MỀM R ĐỂ TIÊN LƯỢNG VÀ ĐÁNH GIÁ BỆNH TIỂU ĐƯỜNG

Trương Thị Hồng Thúy[1], Hà Thị Hằng[1], Đỗ Thị Hồng Nga[1], Bùi Xuân Diện[2]

TÓM TẮT

Đặt vấn đề: Tại các bệnh viện tuyến xã, phường xa trung tâm, việc chẩn đoán bệnh tiểu đường còn chưa kịp thời do khó tiếp cận với các xét nghiệm chẩn đoán. **Mục tiêu:** Bài báo đưa ra giải pháp ứng dụng mô hình hồi quy Logistic kết hợp với các thư viện tích hợp trên phần mềm R để lựa chọn mô hình tối ưu trong tiên lượng và đánh giá tiểu đường. Mục đích để tiết kiệm được chi phí khám bệnh cũng như kịp thời chẩn đoán cho các bệnh nhân ở các tuyến xã, phường xa trung tâm không có điều kiện để thực hiện xét nghiệm. **Địa điểm và phương pháp nghiên cứu:** Nghiên cứu được thực hiện tại trường Đại học Y dược Thái Nguyên bằng phương pháp thực nghiệm. Tác giả sử dụng phần mềm R chạy mô hình hồi quy Logistic trên bộ số liệu bệnh tiểu đường có sẵn. **Kết quả:** Mô hình có tính phân loại rất tốt với độ nhạy đạt 80.60% và độ đặc hiệu là 93.26%.

Từ khóa: Bệnh tiểu đường, mô hình hồi quy Logistic, phần mềm R, mô hình tối ưu, tiên lượng.

SUMMARY

RESEARCH ON BUILDING LOGICSTIC REGRESSION MODEL BY R SOFTWARE TO DIABETES EVALUATE AND PREDICTION

Problem: In commune and ward hospitals far from the centre, diabetes diagnosis is still not on time because it is difficult to access to test. **Objectice:** The paper proposes a solution applied Logistic regression model combined with integrated libraries on R software to select the optimal model for diabetes prediction and assessment which will save costs and promptly diagnose for patients in communes and wards far from the centre that cannot afford testing. **Setting and method:** The research is performed in Thai Nguyen university of medicine and pharmacy by experiment. The authors uses R software to run Logistic regression model on available diabetes data set. **Results:** The results show that the model has a very good classification with the sensitivity reaching 80.60% and the specificity reaching 93.26%.

Keywords: Diabetes, Logistic regression model, R software, optimal model, prediction.

I. ĐẶT VẤN ĐỀ

Bệnh tiểu đường hay còn gọi là đái tháo đường đang trở thành đại dịch nguy hiểm trên thế giới cũng như ở Việt Nam do sự gia tăng nhanh số lượng người mắc, hậu quả nặng nề của bệnh đối với sức khỏe và tuổi thọ, gánh nặng chi phí bệnh cho bệnh nhân và xã hội. Theo kết quả điều tra quốc gia yếu tố nguy cơ bệnh không lây nhiễm năm 2015 (STEP 2015), tỷ lệ người ở độ tuổi từ 18-69 có rối loạn đường huyết khi đói là 3,6% và 4,1% mắc tăng đường huyết (glucose huyết tương tĩnh mạch ≥7,0 mmol) hoặc đang dùng thuốc điều trị đái tháo đường. Những gánh nặng trên đang tạo ra áp lực ngày càng tăng lên hệ thống y tế nước ta. Tuy nhiên, cho đến nay người mắc bệnh đái tháo đường chỉ được quản lý theo dõi ở các cơ sở y tế từ tuyến huyện trở lên [1].

Bệnh tiểu đường sẽ để lại các biến chứng nghiêm trọng nếu không được chẩn đoán và điều trị kịp thời. Để chẩn đoán bệnh tiểu đường cần dựa vào các xét nghiệm cận lâm sàng như: xét nghiệm đường huyết lúc đói, xét nghiệm đường huyết ngẫu nhiên, nghiệm pháp dung nạp glucose đường uống và xét nghiệm HbA1C [1]. Các xét nghiệm này không phải quá tốn kém nhưng tại các bệnh viện tuyến xã, phường xa trung tâm thì điều này gây khó khăn cho bệnh nhân.

Giải pháp đưa ra là ứng dụng các mô hình toán học để tiên lượng và đánh giá tiểu đường trên các phần mềm

1. Trường Đại học Y Dược Thái Nguyên
Tác giả chính Trương Thị Hồng Thuý, SĐT: 0888020959 - Email: truongthihongthuy@tump.edu.vn
2. Trường Đại học Nội vụ Hà Nội

Ngày nhận bài: 07/01/2021 | Ngày phản biện: 25/01/2021 | Ngày duyệt đăng: 06/02/2021

thống kê như R sẽ tiết kiệm được chi phí cũng như kịp thời chẩn đoán cho các bệnh nhân ở các tuyến xã, phường xa trung tâm không có điều kiện để thực hiện xét nghiệm.

II. ĐỐI TƯỢNG VÀ PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Đối tượng nghiên cứu

Mô hình hồi quy Logistic: Hồi quy Logistic là một phương pháp hồi quy thông dụng để xem xét mối liên quan giữa biến độc lập (biến định lượng hoặc biến phân loại) và biến phụ thuộc không phải biến định lượng liên tục mà là biến nhị phân. Trong mô hình hồi quy Logistic có sự nghịch đảo của hàm phân phối xác suất chuẩn hóa [2 - 4]. Mô hình hồi quy Logistic được ứng dụng nhiều để xác định các yếu tố nguy cơ hay dự đoán khả năng xảy ra một sự kiện hay một bệnh nào đó trong y học [2]. Trong nghiên cứu này, chúng tôi sử dụng mô hình hồi quy Logistic đa biến để xác định các yếu tố ảnh hưởng và dự đoán khả năng mắc bệnh tiểu đường từ những biến độc lập có thể đo lường đơn giản hoặc tra cứu tiền sử bệnh nhân (như tiểu nhiều, khát nhiều, cứng cơ, rụng tóc, béo phì,…). Mô hình hồi quy Logistic đa biến có dạng [2]:

$$\log(\frac{P}{1-P}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$$

Trong đó: P là xác suất bệnh nhân mắc bệnh tiểu đường (nhận giá trị là 1); 1-P là xác suất bệnh nhân không mắc bệnh tiểu đường (nhận giá trị 0)

Phần mềm R: Phần mềm R là một ngôn ngữ lập trình phục vụ cho việc tính toán và đồ họa thống kê. Phần mềm này hỗ trợ rất nhiều công cụ để phân tích và xử lý số liệu. R có nền tảng lập trình hướng đối tượng mạnh hơn đa số các phần mềm thống kê khác. Bởi R sử dụng các thư viện tích hợp với các gói mở do người dùng đưa lên và sử dụng đồ thị chất lượng cao [5].

Bộ số liệu tiểu đường: Trong nghiên cứu này, tác giả sử dụng mô hình hồi quy Logistic để lựa chọn ra các đặc trưng quan trọng để xây dựng mô hình đánh giá tiên lượng mắc bệnh tiểu đường với dữ liệu mẫu được lấy từ ngân hàng dữ liệu UCI - Machine learning repository [6] được tải từ địa chỉ:

https://archive.ics.uci.edu/ml/machine-learning-databases/00529/

2.2. Địa điểm và thời gian nghiên cứu

Nghiên cứu được thực hiện tại Trường Đại học Y dược, Đại học Thái Nguyên từ tháng 3 đến tháng 10 năm 2020

2.3. Phương pháp nghiên cứu

2.3.1. Phương pháp và thiết kế nghiên cứu: Nghiên cứu ứng dụng của phần mềm R trong thống kê y học.

2.3.2. Mẫu nghiên cứu

Lấy toàn bộ bộ số liệu tiểu đường từ ngân hàng dữ liệu UCI - Machine learning repository. Bộ số liệu này gồm 520 bản ghi và 17 biến. Cụ thể gồm các biến như sau: Age (Tuổi); Gender (Giới tính); Polyuria (Tiểu nhiều); Polydipsia (Khát nhiều); Sudden weight loss (giảm cân đột ngột); Weakness (yếu); Polyphagia (chứng ăn nhiều); Genital thrush (Tưa miệng); visual blurring (mắt mờ); Itching (Ngứa); Irritability (Cáu gắt); delayed healing (chậm lành vết thương); partial paresis (liệt một phần); muscle stiffness (cứng cơ); Alopecia (rụng tóc từng mảng); Obesity (béo phì) và 1 biến đầu ra (diabetes) mang 2 giá trị Positive (mắc bệnh tiểu đường) và Negative (Không mắc tiểu đường).

2.4. Quản lý và phân tích số liệu

Sử dụng phần mềm R với các thư viện tích hợp để phân tích và xử lý số liệu. Bao gồm các thư viện: Bao gồm các thư viện: UsingR, ggplot2, psych, DescTools, visreg, BMA, epiDisplay, rms, caret, pROC, DynNom và shiny. Kết quả trình bày dưới dạng các bảng và hình ảnh.

III. KẾT QUẢ NGHIÊN CỨU

3.1. Mô tả các biến

Sử dụng phần mềm R để mô tả các biến quan sát trên tập dữ liệu. Sau khi chạy câu lệnh, bảng mô tả dữ liệu được thể hiện ở hình sau:

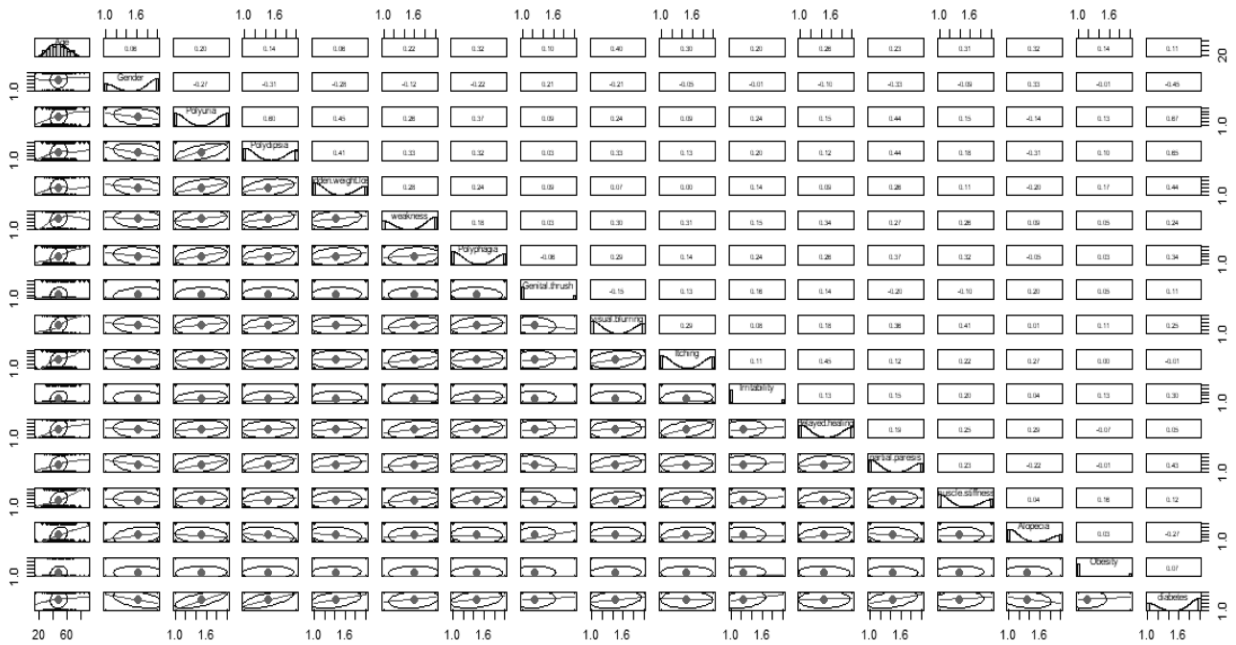
Hình 1. Bảng mô tả các đặc trưng trong tập dữ liệu

```
      Age           Gender     Polyuria  Polydipsia sudden.weight.loss weakness  Polyphagia
 Min.   :16.00   Female:192   No :262   No :287    No :303             No :215   No :283
 1st Qu.:39.00   Male  :328   Yes:258   Yes:233    Yes:217             Yes:305   Yes:237
 Median :47.50
 Mean   :48.03
 3rd Qu.:57.00
 Max.   :90.00
 Genital.thrush visual.blurring Itching   Irritability delayed.healing partial.paresis muscle.stiffness
 No :404        No :287         No :267   No :394      No :281         No :296         No :325
 Yes:116        Yes:233         Yes:253   Yes:126      Yes:239         Yes:224         Yes:195


 Alopecia  Obesity       diabetes
 No :341   No :432   Negative:200
 Yes:179   Yes: 88   Positive:320
```

Bên cạnh đó tác giả cũng xem xét liệu các biến có mối liên quan chặt chẽ với nhau hay không để tránh hiện tượng đa cộng tuyến khi đưa vào mô hình hồi quy bằng cách sử dụng biểu đồ mô tả mối liên quan giữa các biến trong mô hình. Kết quả thể hiện trong hình 2.

Hình 2. Biểu đồ mô tả mối liên quan giữa các biến trong tập dữ liệu



3.2. Lựa chọn đặc trưng để chọn mô hình tối ưu

Đầu tiên, chạy mô hình hồi quy với tất cả các đặc trưng đầu vào của bộ số liệu ta thu được kết quả mô hình hồi quy Logistic như sau:

Hình 3. Kết quả đánh giá mô hình hồi quy với tất cả các biến đầu vào

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	520	LR chi2	445.47	R2	0.782	C	0.963
Negative	200	d.f.	15	g	4.615	Dxy	0.926
Positive	320	Pr(> chi2)	<0.0001	gr	101.012	gamma	0.926
max |deriv|	1e-08			gp	0.439	tau-a	0.439
				Brier	0.070		

	Coef	S.E.	Wald Z	Pr(>|Z|)
Intercept	1.5616	0.8230	1.90	0.0578
Age	-0.0313	0.0191	-1.64	0.1004
Gender=Male	-3.0383	0.4437	-6.85	<0.0001
Polyuria=Yes	3.8251	0.4644	8.24	<0.0001
sudden.weight.loss=Yes	0.9758	0.4012	2.43	0.0150
weakness=Yes	0.9102	0.4021	2.26	0.0236
Polyphagia=Yes	0.6018	0.4138	1.45	0.1458
Genital.thrush=Yes	1.8027	0.4465	4.04	<0.0001
visual.blurring=Yes	1.3003	0.5118	2.54	0.0111
Itching=Yes	-1.4946	0.4529	-3.30	0.0010
Irritability=Yes	2.2454	0.5001	4.49	<0.0001
delayed.healing=Yes	-0.9876	0.4449	-2.22	0.0264
partial.paresis=Yes	1.0258	0.4091	2.51	0.0122
muscle.stiffness=Yes	-0.2622	0.4373	-0.60	0.5487
Alopecia=Yes	-0.5637	0.4671	-1.21	0.2275
Obesity=Yes	-0.2101	0.4557	-0.46	0.6447

Mô hình này cho kết quả là tổng bình phương của các phần dư (Residual Sum of Squares) trong mô hình ở mức R^2=0.782 tức là mô hình có thể giải thích được khoảng 78.2% khác biệt về nguy cơ mắc bệnh tiểu đường.

Tuy nhiên, quan sát giá trị của các biến trong mô hình hồi quy thấy rằng, mô hình hồi quy tương đối phức tạp và có rất nhiều biến không có ý nghĩa trong mô hình (p_value>0.05). Liệu có thể tìm một mô hình sử dụng ít biến đầu vào hơn mà có thể giải thích được khác biệt về nguy cơ mắc bệnh tiểu đường tương đương hoặc tốt hơn.

Sử dụng công cụ Bayesian Model Averageing (BMA) trên phần mềm R để tìm mô hình tối ưu. Sau khi chạy thuật toán thì có mô hình được lựa chọn thu được 105 mô hình. Trong đó, có 5 mô hình tốt nhất được lựa chọn. Mỗi mô hình BMA sẽ báo cáo hệ số hồi quy của mỗi biến tiên lượng, hệ số xác định R^2 và xác suất hậu nghiệm.

Hình 4. Kết quả chạy thuật toán BMA để lựa chọn các mô hình tối ưu nhất

Best 5 models (cumulative posterior probability = 0.2082):

	p!=0	EV	SD	model 1	model 2	model 3	model 4	model 5
Intercept	100	0.849456	0.66473	0.5925	0.6779	0.5235	0.7202	0.7774
Age	15.9	-0.005758	0.01513	.	.	.	.	.
Gender	100.0							
.Male		-3.082734	0.42672	-3.2060	-2.9827	-3.0891	-3.1016	-3.2848
Polyuria	100.0							
.Yes		3.767369	0.44997	3.8029	3.6971	3.6490	3.6851	3.9569
sudden.weight.loss	78.9							
.Yes		0.904674	0.57823	0.9844	1.1608	1.2574	1.2243	.
weakness	43.6							
.Yes		0.432996	0.55848	0.9498	.	.	.	1.1910
Polyphagia	9.7							
.Yes		0.062986	0.22617	.	.	.	.	.
Genital.thrush	94.6							
.Yes		1.369749	0.55423	1.3886	1.4128	1.6071	1.2772	1.5067
visual.blurring	53.5							
.Yes		0.644774	0.69409	.	1.3127	1.0452	.	.
Itching	93.8							
.Yes		-1.297677	0.54867	-1.2417	-1.4224	-1.4100	-1.0008	-1.3217
Irritability	100.0							
.Yes		2.199678	0.48063	2.1830	2.2412	2.3178	2.3240	2.1056
delayed.healing	55.7							
.Yes		-0.619196	0.63872	-1.2103	.	-1.0927	-0.9629	-1.2162
partial.paresis	69.7							
.Yes		0.767240	0.60550	1.2234	.	0.9861	1.2169	1.2353
muscle.stiffness	0.0							
.Yes		0.000000	0.00000	.	.	.	.	.
Alopecia	38.2							
.Yes		-0.386852	0.55703	.	-1.0879	.	.	.
Obesity	1.6							
.Yes		-0.007413	0.08033	.	.	.	.	.
nVar				9	8	9	8	8
BIC				-2927.4764	-2927.2613	-2926.9100	-2926.9000	-2926.7735
post prob				0.051	0.046	0.038	0.038	0.036

Chạy mô hình hồi quy tối ưu sử dụng 9 biến đặc trưng theo mô hình tốt nhất đã chọn theo thuật toán bên trên là: Gender (Giới tính); Polyuria (Tiểu nhiều); Sudden weight loss (giảm cân đột ngột); Weakness (yếu); Genital thrush (Tưa miệng); Itching (Ngứa); Irritability (Cáu gắt); delayed healing (chậm lành vết thương); partial paresis (liệt một phần). Kết quả thu được như sau:

Hình 5. Kết quả chạy mô hình tối ưu

```
                      Model Likelihood     Discrimination    Rank Discrim.
                         Ratio Test           Indexes           Indexes
Obs           520    LR chi2     430.95    R2        0.765    C         0.957
 Negative     200    d.f.             9    g         4.311    Dxy       0.915
 Positive     320    Pr(> chi2) <0.0001    gr       74.533    gamma     0.918
max |deriv| 8e-12                          gp        0.435    tau-a     0.434
                                           Brier     0.075

                        Coef    S.E.   Wald Z Pr(>|Z|)
Intercept               0.5925 0.3941  1.50  0.1327
Gender=Male            -3.2060 0.4162 -7.70  <0.0001
Polyuria=Yes            3.8029 0.4307  8.83  <0.0001
sudden.weight.loss=Yes  0.9844 0.3752  2.62  0.0087
weakness=Yes            0.9498 0.3672  2.59  0.0097
Genital.thrush=Yes      1.3886 0.4019  3.45  0.0006
Itching=Yes            -1.2417 0.3870 -3.21  0.0013
Irritability=Yes        2.1830 0.4502  4.85  <0.0001
delayed.healing=Yes    -1.2103 0.3947 -3.07  0.0022
partial.paresis=Yes     1.2234 0.3733  3.28  0.0010
```

3.3. Huấn luyện và đánh giá mô hình

Để huấn luyện và đánh giá mô hình vừa chọn, tác giả thực hiện phương pháp chia tập dữ liệu thành 2 tập: tập huấn luyện (70% dữ liệu mẫu) và tập kiểm tra (30% dữ liệu mẫu). Sau khi chạy mô hình trên tập huấn luyện thu được kết quả:

Hình 6. Kết quả huấn luyện trên mô hình tối ưu

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.92366  -0.38966   0.04446   0.25943   2.93206

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.9348     0.5124   1.824 0.068125 .
GenderMale           -3.4747     0.5281  -6.579 4.73e-11 ***
PolyuriaYes           4.0556     0.5423   7.478 7.55e-14 ***
sudden.weight.lossYes 0.8494     0.4717   1.801 0.071751 .
weaknessYes           1.0097     0.4216   2.395 0.016620 *
Genital.thrushYes     1.4647     0.4813   3.043 0.002342 **
ItchingYes           -1.7449     0.4914  -3.551 0.000384 ***
IrritabilityYes       2.3154     0.5545   4.175 2.97e-05 ***
delayed.healingYes   -1.0927     0.4684  -2.333 0.019651 *
partial.paresisYes    1.1621     0.4489   2.589 0.009624 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 485.05  on 363  degrees of freedom
Residual deviance: 182.46  on 354  degrees of freedom
AIC: 202.46
```

Sử dụng ma trận hỗn hợp để đánh giá mô hình tập kiểm tra, cho kết quả như sau:

Hình 7. Ma trận hỗn hợp của 2 mô hình

```
Confusion Matrix and Statistics

          Reference
Prediction Negative Positive
  Negative       51        9
  Positive        8       88

               Accuracy : 0.891
                 95% CI : (0.8313, 0.9352)
    No Information Rate : 0.6218
    P-Value [Acc > NIR] : 3.738e-14

                  Kappa : 0.7691

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8644
            Specificity : 0.9072
```

Mô hình ban đầu (16 biến đầu vào)

```
Confusion Matrix and Statistics

          Reference
Prediction Negative Positive
  Negative       54        6
  Positive       13       83

               Accuracy : 0.8782
                 95% CI : (0.8164, 0.9251)
    No Information Rate : 0.5705
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7482

 Mcnemar's Test P-Value : 0.1687

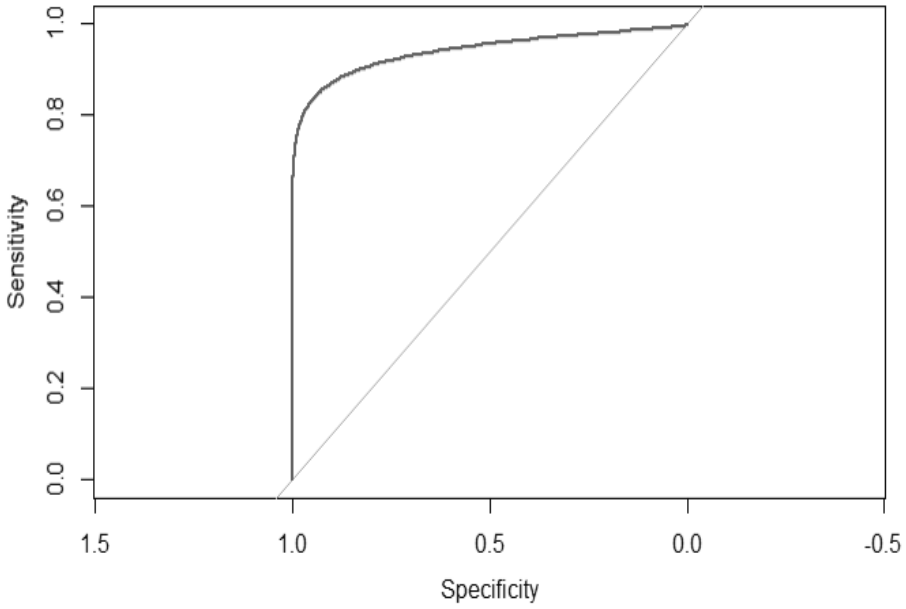
            Sensitivity : 0.8060
            Specificity : 0.9326
```

Mô hình tối ưu (9 biến đầu vào)

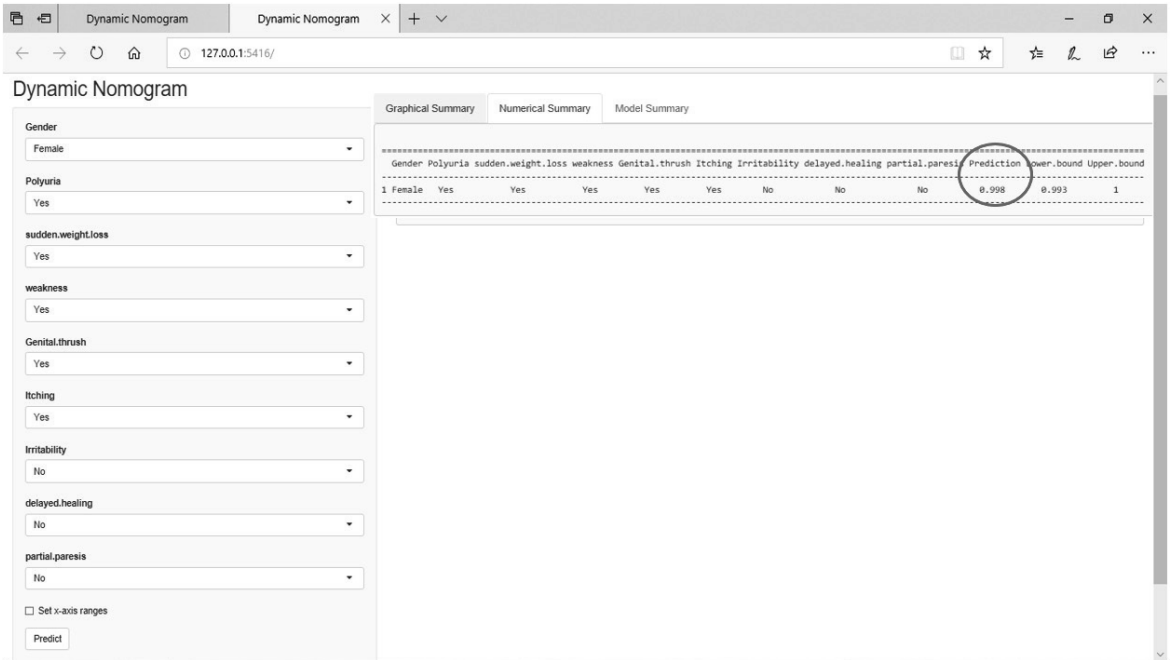
Để thể hiện kết quả rõ hơn, tác giả đã vẽ đường cong ROC (Receiver Operating Characteristics) để đánh giá chất lượng cường độ phân biệt của mô hình thử nghiệm sử dụng số liệu có biểu thị độ nhạy và độ đặc hiệu. Đường cong này là tập hợp những gia điểm có hoành độ là tỷ lệ âm tính thật và tung độ là độ nhạy của từng giá trị thử nghiệm.

Diện tích dưới đường cong trong mô hình vừa chạy bên trên là: 0.9484. Như vậy, việc xây dựng mô hình toán học với các đặc trưng được lựa chọn có ý nghĩa trong việc đánh giá tiên lượng mắc bệnh tiểu đường khá tốt. Tuy nhiên, với những người dùng không chuyên về toán học thì mô hình có ý nghĩa thực tế không cao. Tác giả đã sử dụng thêm công cụ Dynamic Nomogram trên R để chạy chương trình chẩn đoán với các bệnh nhân mới. Người dùng chỉ cần nhập các thông tin đầu vào của bệnh nhân, hệ thống sẽ tính toán xác suất mắc bệnh dựa trên mô hình hồi quy đã xây dựng.

Hình 8. Đường cong ROC



Hình 11: Kết quả chạy ứng dụng Dynom để tiên lượng mắc bệnh tiểu đường cho bệnh nhân mới



IV. BÀN LUẬN

4.1. Lựa chọn mô hình tối ưu

Mô hình hồi quy Logistic có nhiều ưu điểm trong tiên lượng, đánh giá biến kết quả từ các biến đầu vào [2 - 4]. Khi xây dựng mô hình hồi quy Logistic để tiên lượng xác suất mắc bệnh tiểu đường, việc sử dụng tất cả các biến đầu vào (16 biến) trong tập dữ liệu để đưa vào mô hình làm cho phương trình hồi quy rất phức tạp. Ngoài ra, có rất nhiều biến đặc trưng không có ý nghĩa trong mô hình (với $p > 0,05$). Vì vậy tác giả đã sử dụng thuật toán BMA để lựa chọn mô hình tối ưu nhất trong 105 mô hình được lựa chọn. Dựa trên kết quả của xác suất hậu định và chỉ số BIC, tác giả đã chọn mô hình tối ưu nhất gồm 9 biến đặc trưng đầu vào bao gồm Gender (Giới tính); Polyuria (Tiểu nhiều); Sudden weight loss (giảm cân đột ngột); Weakness (yếu); Genital thrush (Tưa miệng); Itching (Ngứa); Irritability (Cáu gắt); delayed healing (chậm lành vết thương); partial paresis (liệt một phần). Sau khi chạy mô hình hồi quy thì phương trình hồi quy được xây dựng thể hiện trong hình 5.

4.2. Đánh giá mô hình tối ưu

Việc sử dụng 9 biến đặc trưng đầu vào để xây dựng mô hình hồi quy cho kết quả R^2=76,5%, tức là mô hình có thể giải thích được khoảng 76,5% khác biệt về nguy cơ mắc bệnh tiểu đường. Theo các tài liệu [2, 7], mô hình được lựa chọn là mô hình có hệ số xác định R2 lớn nhất. Tuy nhiên, mô hình này có hệ số xác định R^2 là thấp hơn so với mô hình ban đầu có R^2=78,2% nhưng thấp hơn không đáng kể và sử dụng ít hơn rất nhiều số biến đặc trưng đầu vào làm cho mô hình hồi quy đơn giản hơn. Ngoài ra, hầu hết tất cả các đặc trưng trong mô hình được lựa chọn đều có liên quan đến bệnh tiểu đường với $p < 0,01$.

Bên cạnh đó tác giả cũng sử dụng phương pháp học máy để huấn luyện và đánh giá mô hình bằng cách chia tập dữ liệu ra làm 2 tập con để huấn luyện (70% dữ liệu mẫu) và đánh giá (30% dữ liệu mẫu). Kết quả trong hình 6 cho thấy khi đưa 9 biến đầu vào đã được lựa chọn vào mô hình thì chúng có ảnh hưởng đến việc chẩn đoán mắc bệnh tiểu đường đáng kể, làm cho giảm độ khác biệt về giá trị quan sát và ước tính kết quả từ 485.05 xuống 182.46 trên tập dữ liệu huấn luyện. Kết quả trên hình 7 sử dụng ma trận hỗn hợp để đánh giá trên tập kiểm tra cho độ nhạy (sencitivity) đạt 80.6% và độ đặc hiệu (specificity) đạt 93,26% (Sử dụng cách làm tương tự mô hình ban đầu với 16 đặc trưng thì thu được độ nhạy là 86,44% và độ đặc hiệu là 90,72%). Như vậy, có thể thấy mô hình sử dụng 9 biến đặc trưng đã được lựa chọn về cơ bản tương đương với mô hình ban đầu gồm 16 đặc trưng trong việc tiên lượng đánh giá nguy cơ mắc tiểu đường.

Ngoài ra, tác giả đã vẽ đường cong ROC (Receiver Operating Characteristics) để đánh giá chất lượng cường độ phân biệt của mô hình thử nghiệm sử dụng số liệu có biểu thị độ nhạy và độ đặc hiệu. Kết quả hiển thị trong hình 8 cho diện tích dưới đường cong trong mô hình tối

ưu là 0,9484. Do đó, mô hình có khả năng phân định bệnh nhân có bị mắc tiểu đường hay không rất tốt.

4.3. Ứng dụng tiên lượng để chẩn đoán nguy cơ mắc tiểu đường cho bệnh nhân mới

Việc xây dựng mô hình toán học với các đặc trưng được lựa chọn có ý nghĩa trong việc đánh giá tiên lượng mắc bệnh tiểu đường khá tốt. Tuy nhiên, với những người dùng không chuyên về toán học thì mô hình có ý nghĩa thực tế không cao. Tác giả đã sử dụng thêm công cụ Dynamic Nomogram tích hợp trên R để chạy chương trình chẩn đoán với các bệnh nhân mới. Người dùng chỉ cần nhập các thông tin đầu vào của bệnh nhân, hệ thống sẽ tính toán xác suất mắc bệnh dựa trên mô hình hồi quy đã xây dựng. Trong hình 9 minh họa với các đặc trưng được lựa chọn thì xác suất mắc bệnh tiểu đường của bệnh nhân là 99,8%.

Mặc dù mô hình cho kết quả rất tốt, có độ tin cậy cao trên tập dữ liệu mẫu nhưng theo các chuyên gia, việc ứng dụng mô hình này trong thực tế cần nghiên cứu thêm và cần thử nghiệm so sánh giữa việc chẩn đoán của phần mềm và của các bác sĩ. Kết quả nghiên cứu này chứng minh khả năng ứng dụng của các công cụ toán học để giải quyết các bài toán thực tế như bài toán hỗ trợ ra quyết định trong y học.

V. KẾT LUẬN

Trong khoảng 5 năm trở lại đây, phần mềm thống kê R đã được ứng dụng và phát triển rộng rãi trong nghiên cứu. Việc ứng dụng phần mềm R nói chung và mô hình hồi quy Logistic trong hỗ trợ ra quyết định chẩn đoán bệnh sẽ hỗ trợ được các bác sĩ ở các tuyến xã, phường không có điều kiện để thực hiện xét nghiệm kịp thời tiên lượng bệnh nhân có hay không bị mắc bệnh.

Bằng việc ứng dụng mô hình hồi quy logistic, tác giả đã lựa chọn các đặc trưng tối ưu nhất để xây dựng mô hình, sử dụng ma trận hỗn hợp và đường cong ROC để đánh giá mô hình cho kết quả rất tốt với độ nhạy đạt 80,60%, độ đặc hiệu đạt 93,26%, diện tích dưới đường cong đạt 0,9484. Bên cạnh đó, tác giả còn sử dụng ứng dụng Dynamic Nomogram để tính toán xác suất mắc bệnh của các bệnh nhân mới, người dùng chỉ cần lựa chọn các đặc trưng đầu vào bằng các thao tác nhấn chuột, kết quả sẽ hiển thị xác suất mắc bệnh được tính dựa trên phương trình hồi quy đã xây dựng.

TÀI LIỆU THAM KHẢO

1. Bộ Y tế (2017). "Hướng dẫn chẩn đoán và điều trị đái tháo đường típ 2".

2. Nguyễn Văn Tuấn (2018). "Phân tích dữ liệu với R Hỏi và đáp". Nhà xuất bản Tổng hợp thành phố Hồ Chí Minh.

3. Nguyễn Thanh Danh, Đậu Văn Ngọ, Tạ Quốc Dũng (2017). Ứng dụng phương pháp hồi quy Logistic xác định tổ hợp tối ưu các yếu tố ảnh hưởng và xây dựng bản đồ tai biến trượt lở đất huyện Khánh Vĩnh, tỉnh Khánh Hòa. *Tạp chí Phát triển Khoa học và C*ông nghệ, 2017.

4. http://r-statistics.co/Logistic-Regression-With-R.html.

5. Paul T (2011). "R codebook". O'Reilly.

6. https://archive.ics.uci.edu/ml/machine-learning-databases/00529/

7. Max K, Kjell J (2013). "Applied Predictive Modeling". Springer.