1.Choose one of the problems to apply Data Science Methodology

Tiến hành chọn chủ đề 3:

“Data from the National Heart Association from 2012 shows 65% of people with diabetes will die from some sort of heart disease or stroke and kidney disease. Identify who has diabetes will lead to a better understanding of these complications and the common comorbidities that diabetics suffer. The challenge to you is using medical record data to improve public health by determining patients diagnosed with Type 2 Diabetes Mellitus.”

Dịch là: “Dữ liệu từ Hiệp hội Tim mạch Quốc gia từ năm 2012 cho thấy 65% ​​những người bị bệnh tiểu đường sẽ chết vì một số loại bệnh tim hoặc đột quỵ và bệnh thận. Xác định người mắc bệnh tiểu đường sẽ dẫn đến sự hiểu biết tốt hơn về các biến chứng và bệnh đi kèm thường gặp mà bệnh nhân tiểu đường phải chịu. Các thách thức với bạn là sử dụng dữ liệu hồ sơ y tế để cải thiện sức khỏe cộng đồng bằng cách xác định bệnh nhân được chẩn đoán mắc bệnh tiểu đường loại 2”

2.

Trước tiên đây là một đề bài rất thú vị và mang tính thực tiễn rất cao. Khi đời sống ngày càng cao thì vấn đề về sức khỏe hay nhiều bệnh kéo theo do lối sống hay sinh hoạt đa dạng cũng tăng theo, vấn đề chăm sóc sức khỏe y tế đã trở nên cấp thiết hơn bao giờ hết. Tiểu đường được cho là căn bệnh thường gặp trong cuộc sống hiện đại, do lượng glucose trong máu vượt ngưỡng quy định(>11,1mmol/l). Bệnh tiểu đường do nhiều nguyên nhân, căn cứ vào các nguyên nhân người ta chia thành bốn loại: tiểu đường tuýp 1, tiểu đường tuýp 2, tiểu đường thứ phát và tiểu đường thai kỳ. Trong đó tiểu đường tuýp 2 là căn bệnh chủ yếu do rối loạn lối sống. “Đây là bệnh xảy ra do chế độ ăn uống với hàm lượng calo cao, nhiều chất béo, thiếu vận động và do sự giảm tiết insulin hoặc tác dụng của insulin không hiệu quả khiến lượng đường trong máu tăng cao. Ở giai đoạn đầu của bệnh thường rất ít triệu chứng cơ năng và cũng có trường hợp lượng đường trong máu cao kéo dài. Khi tình trạng tăng đường huyết xảy ra, glucose còn lại trong máu tác động xấu đến tuyến tụy, làm giảm lượng tiết insulin và giảm hiệu quả của insulin đến các mô như gan và cơ bắp hoặc gây ra tình trạng kháng insulin. Kết quả là, các mạch máu và dây thần kinh bị tổn thương và dẫn đến các biến chứng khác nhau xảy ra.”( Theo <https://kienthuctieuduong.vn/su-khac-nhau-giua-tieu-duong-tuyp-1-2-la-gi/>). Như vậy điểm quan trọng khác biệt ở tiểu đường tuýp 2 là kiểm soát được các biến chứng ngoài việc kiểm soát được lượng đường trong máu. Theo thống kê thì các biến chứng của tiểu đường tuýp 2 thường gặp như bệnh tim, đột quỵ và suy thận. Đây đều là những biến chứng nguy hiểm là nguyên nhân gây tử vong hàng đầu ở người bị tiểu đường tuýp 2. Việc xác định bệnh nhân bị tiểu đường tuýp 2 chính xác có ý nghĩa sống còn vì xác định bệnh nhân đồng nghĩa với việc chuẩn bị các phương án kiểm soát với các biến chứng nguy hiểm hay xảy ra như đã đề cập trên đây.

Cải thiện sức khỏe cộng đồng luôn là một đề tài thực tiễn, ý nghĩa, thú vị nhưng cũng rất rộng lớn và không kém phần phức tạp và khó khăn. Đầu ra đôi khi mơ hồ không rõ ràng trong khi đầu vào là rất nhiều biến số. Để xác định bệnh nhân nào bị bệnh tiểu đường khi nhập viện với các triệu chứng lâm sàng thường gặp thì chỉ cần dựa vào kinh nghiệm chuẩn đoán của các bác sĩ và kết quả của các xét nghiệm tiểu đường xác định hàm lượng glucose trong máu là đủ. Tuy nhiên khi chúng ta muốn một hệ thống làm được nhiều hơn thế, một hệ thống mang tính cộng đồng cao hơn có khả năng đi xa hơn là chuẩn đoán, đó sẽ là một hệ thống dự báo, hệ thống dự báo ai sẽ có nguy cơ bị tiểu đường dựa trên những thông tin được khảo sát từ người tham gia (bất kỳ ai cũng có thể tham gia không). Nhiệm vụ của chúng ta sẽ là tổ chức một hình thức phù hợp để hệ thống tương tác với người tham gia bằng những câu hỏi khảo sát với mục đích thu thập dữ liệu. Những dữ liệu này có thể là các thông tin về tên, tuổi, tiền sử bệnh nhận, mô tả tình trạng triệu chứng hiện tại hay cụ thể hơn là các chỉ số về chiều cao, cân nặng, huyết áp, lượng đường trong máu..vv. Mô mình mà hệ thống vận hành sẽ sử dụng những dữ liệu này để phân tích và đưa ra kết quả người tham gia có bị bệnh tiểu đường hay không, hoặc người tham gia có nguy cơ sẽ bị tiểu đường hay không. Bằng các kết quả đó những lời khuyên hoặc khuyến cáo tương ứng với từng trường hợp sẽ hệ thống tự động đưa ra. Về hình thức thì mô hình dự án của chúng ta có thể được triển khai dưới dạng một trang web hay một phần mềm hoặc kết hợp cả hai hình thức. Hệ thống có thể kết nối giữa người sử dụng với bệnh viện, các chuyên gia, hoặc bác sĩ nhằm giải đáp chi tiết hơn những thắc mắc về kết quả, hoặc chuẩn bị điều trị đối với những người được cho là mắc bệnh, hoặc là những lời khuyên để cải thiện tình trạng lối sống hiện tại nhằm hướng tới một cơ thể khỏe mạnh hơn giúp không chỉ tránh được bệnh tiểu đường mà còn cả các bệnh khác.

Như vậy mục tiêu của dự án (***Business goals***) sẽ là:

***Dựa trên hồ sơ dữ liệu bệnh nhân xây dựng hệ thống chuẩn đoán dự báo người mắc bệnh tiểu đường tuýp 2 hoặc người có nguy cơ cao sẽ bị tiểu đường tuýp 2.***

Các mục tiêu triển khai cụ thể hơn (***Business Objectives)*** nhằm đảm bảo hệ thống được cụ thể nhất có thể từ hình thức đến nội dung.

**+** Triển khai hệ thống dưới dạng trang Web hoặc phần mềm tương tác với số lượng người sử dụng tối thiểu cần đạt được (lý tưởng sẽ là đến từng người dân).

**+** Nhiệm vụ chính của hệ thống sẽ là dự đoán cho nên tỉ lệ dự đoán chính xác sẽ là yêu cầu cần được quản lý của hệ thống này.

Trong khuôn khổ dự án được tài trợ chúng ta sẽ cân nhắc xem việc định lượng chính xác những mục tiêu nào, những mục tiêu nào chỉ cần trong mức độ định tính. Ví dụ mục tiêu đầu là triển khai đến số lượng người có dùng máy tính hoặc điện thoại thông minh có vẻ đơn giản tuy nhiên muốn phổ cập đến cả các thành phần còn lại có thể sẽ là nhiệm vụ khó khăn. Với mục tiêu thứ hai tỉ lệ dự đoán chính xác ở đây là tỉ lệ biểu kiến. Hệ thống sẽ được đánh giá theo thời gian qua tỉ lệ thực tế mà nó dự đoán đúng. Để làm được điều này một sự duy trì theo dõi chăm sóc người dùng từ khi tham gia đến khi xác nhận bị bệnh tiểu đường và lâu hơn về sau nữa là thực sự cần thiết. Sự phản hồi thường xuyên về tình trạng sức khỏe của người sử dụng chính là điều kiện để cập nhập và tinh chỉnh hệ thống.

3.

***Analytic Approach*** . Dựa trên phân tích yêu cầu của đề án và với mục tiêu đã định hình trên đây ta có thể khẳng định lại rằng phương pháp tiếp cận của dự án sẽ là một mô hình phân tích dự báo (***Predictive model)***

Theo nghiên cứu những đối tượng có nguy cơ mắc bệnh tiểu đường cao hơn người thường là: có tiền sử người thân bị tiểu đường, tiền sử đái tháo đường thai kỳ, tuổi cao, ít hoạt động thể chất, thừa cân, béo phì, tăng huyết áp,..vv. Hồ sơ bệnh án của bệnh nhân sẽ bao gồm rất nhiều các chỉ số các thông tin khác nhau mà ta có thể dựa vào để chuẩn đoán. Ví dụ một bệnh nhân khá cao tuổi ( khoảng 60 chẳng hạn) bị nhập viện vì đau tim, theo thông tin hồ sơ bệnh án quan sát được bệnh nhân có người thân bị tiểu đường thì ta có thể phân loại ngay bệnh nhân vào nhóm có nguy cơ mắc bệnh tiểu đường và tiến hành xét nghiệm tiểu đường. Trong trường hợp này phương pháp phân loại (***Classification***) được sử dụng. Khi sử dụng phương pháp trên chúng ta đã mặc định sử dụng kết quả sẵn có của nghiên cứu rằng bệnh tiểu đường có tính di truyền. Nói cách khác trả lời câu hỏi có hoặc không thì xác suất bệnh nhân trên bị tiểu đường khá là cao. Ngược lại nếu một bệnh nhân khác trẻ tuổi hơn (khoảng 30 tuổi chẳng hạn) cũng nhập viện với triệu chứng đau tim, để khẳng định bệnh nhân có nguy cơ tiểu đường hay không trong khi không rõ tiểu sử người thân của bệnh nhân, ta lại phải cần thêm các dữ liệu khác như bệnh nhân có thừa cân béo phì, thói quen sinh hoạt ra sao? Ngoại trừ yếu tố di truyền có xác suất khẳng định cao thì các yếu tố khác như là chỉ số BMI (Body Mass Index), huyết áp, nồng độ glucose là những biến số được quan tâm. Đây là những biến số được định lượng rõ ràng nên ta có thể nghĩ ngay đến một mô hình hồi quy(***Regression analysis***) để tính được xác suất bệnh nhân bị tiểu đường là bao nhiêu %?. Đầu ra của mô hình (gọi là đại lượng Y) là xác suất bệnh nhân sẽ bị bệnh tiểu đường (từ 0 đến 1) còn đầu vào là các biến số (gọi là đại lượng Xi) khác nhau như chỉ số BMI, huyết áp, nồng độ glucose ..vv. Công thức sẽ được mô tả ngắn gọn như sau: Y=∑ai\*Xi + b. Trong đó ‘ai’ gọi là các trọng số, b là độ lệch. Ta ví dụ có 3 biến X1, X2, X3 lần lượt là các chỉ số glucose, BMI và huyết áp. Với Y=1và nếu các trọng số tương ứng là a1>a2>a3 thì có nghĩa là nguyên nhân gây nên tiểu đường đối với glucose là cao nhất và tương ứng thấp nhất sẽ là huyết áp.

Hệ thống hồi quy tuyến tính sẽ dùng tập dữ liệu học (Traning Data) để tìm ra bộ thông số ai và b. Một người tham gia khảo sát được chuẩn đoán mắc bệnh với xác suất là bao nhiêu chỉ cần nhập các thông số về BMI, huyết áp và nồng độ glucose vào công thức trên sẽ tính được giá trị Y là bao nhiêu. Ví dụ Y=0.9 có nghĩa là xác suất người này mắc bệnh sẽ là 90%, nếu Y=0.5 có nghĩa là người này thuộc nhóm nguy cơ có thể sẽ mắc bệnh, lúc này một cảnh báo tự động được đưa ra để người dùng tự chuẩn bị cho mình các phương án cải thiện lối sống. Như vậy ta có thể kết hợp cả 2 cách tiếp cận phân tích trong một mô hình dự báo để chuẩn đoán người sử dụng mắc bệnh hay không.

***Data requirements and data collection:*** Để xây dựng mô hình dự báo của chúng ta sẽ cần dữ liệu bao gồm các thông tin về bệnh nhân như là: giới tính, tuổi, cân nặng, chiều cao (có thể gộp lại thành chỉ số BMI), nồng độ glucose, chỉ số huyết áp hay thông tin về tiền sử bệnh nhân (người nhà mắc bệnh tiểu đường hay không) và tình trạng bệnh hiện tại (bao gồm tiểu đường và cả các bệnh khác đặc biệt là các bệnh kèm theo như tim mạch, đột quỵ hay suy thận). Trong đó giới tính là dạng “text”, tiền sử bệnh nhân là loại “bool” (có hay không người nhà bị bệnh), tuổi là số nguyên và các chỉ số còn lại là số thực. Riêng tình trạng bệnh là thông tin của biến đầu ra ta có thể quy định bằng các mức độ từ 1 đến 5 tùy theo cấp độ nặng, nhẹ khác nhau. Mỗi một bản ghi thông tin về người bệnh sẽ bao gồm các thông tin trên và chúng ta có thể cần khoảng 1000 bản ghi này. Mỗi bản ghi gồm các cột là các chỉ số đã nêu trên. Sau khi xác định được định dạng dữ liệu cho mô hình ta tiến hành xác nhận về phương án thu thập dữ liệu (***collection***) cho mô hình. Nơi thu thập dữ liệu là ở trong một bệnh viện có điều trị bệnh tiểu đường. Nguồn dữ liệu sẽ được lấy từ các hồ sơ bệnh án của các bệnh nhân trong bệnh viện , bao gồm hồ sơ của những bệnh nhân tiểu đường với mức độ nặng nhẹ khác nhau (ví dụ về hồ sơ của một bệnh nhân bị bệnh tiểu đường <https://yhoctonghop.vn/benh-an-dai-thao-duong>), những hồ sơ của những bệnh nhân được xác định không mắc bệnh tiểu đường và cả những hồ sơ của những người khám sức khỏe được xác nhận là bình thường khác . Trong bước này một số công cụ dùng để thu thập, xử lý dữ liệu một cách tự động có thể được các lập trình viên và quản trị viên dữ liệu phối hợp triển khai. Ví dụ như đọc hồ sơ bệnh án của bệnh nhân và lưu vào bản ghi được thiết kế trước đó, vv.

***Data Understanding and Data Preparation:*** Chúng ta cần phải hiểu được dữ liệu bắt đầu từ tìm hiểu diễn biến của các biến số, chỉ số như đã đề cập ở trên. Bước này ta có thể dùng các công cụ thống kê như Excel để tính và vẽ biểu đồ biểu diễn giúp thể hiện trực quan hơn các đại lượng qua các gia trị như min, max, trung bình và độ lệch chuẩn. Ví dụ để kiểm tra xem độ tuổi trẻ nhất bị mắc tiểu đường là bao nhiêu, hay có phải là tỉ lệ mắc bệnh tăng dần theo độ tuổi hay không? Hàm lượng glucose cao nhất được ghi nhận là bao nhiêu? Huyết áp trung bình của nhóm người mắc bệnh tiểu đường so với nhóm bệnh nhân khác hay nhóm người bình thường? Các bệnh đi kèm tiểu đường như tim mạch, đột quỵ hay suy thận chiếm bao nhiêu % so với phần còn lại. Hoặc ta có thể dùng hệ số tương quan để kiểm tra mối quan hệ của từng cặp biến ra sao? Đối với bệnh nhân tiểu đường thì hệ số tương quan giữa chỉ số BMI và hàm lượng glucose là bao nhiêu, cao hay thấp? Hệ số càng tiến tới 1 có nghĩa là 2 chỉ số tương quan mạnh, phụ thuộc lẫn nhau. Nếu hệ số là rất cao thì ta có thể bỏ bớt đi một biến số, chỉ cần giữ lại một đại diện mà thôi. Mục đích của bước hiểu dữ liệu không những giúp ta nắm được dữ liệu nào có ý nghĩa, xu hướng tác động đối với tình trạng bệnh nhân ra sao mà còn kiểm tra được sự đầy đủ và chính xác của việc thu thập dữ liệu để chuẩn bị cho bước chuẩn bị dữ liệu (***Data preparation***). Một yêu cầu về tăng số lượng bản ghi hoặc loại bỏ các chỉ số có thể được đưa ra trong bước này. Trong bước chuẩn bị dữ liệu ta cần thiết phải chuẩn hóa lại dữ liệu để có thể dùng công thức tính toán hay biểu diễn đồ thị, ví dụ như chỉ số huyết áp được ghi là 120/81mm/Hg thì ta cần cân nhắc dùng chỉ số nào hay là quy đổi ra một số tương đương khác. Chẳng hạn chỉ số BMI được thừa nhận sử dụng rộng rãi vốn dĩ cũng được quy đổi từ hai đại lượng là chiều cao và cân nặng theo công thức: chỉ số BMI = cân nặng / chiều cao bình phương. Đối với các giá trị bị thiếu chúng ta cân nhắc có cần loại bỏ bản ghi này đi không hay là tìm cách thích hợp để đánh dấu việc bị thiếu dữ liệu này, ví dụ về độ tuổi thường là từ 0 đến 100 nếu bị thiếu thì ta có thể ghi là 999. Nếu như cần khảo sát giữa chỉ số BMI và huyết áp ta vẫn có thể dùng các dữ liệu này một cách bình thường. Như vậy với việc xử lý dữ liệu bị thiếu và chuyển đổi dữ liệu chúng ta đã có thể sẵn sàng cho bước tiếp theo là lập mô hình của hệ thống dự báo.