PHƯƠNG PHÁP DỰ ĐOÁN TỶ LỆ TỰ TỬ TẠI CÁC NƯỚC TRÊN THẾ GIỚI

Method of predicting the suicide rate of countries in the world

Lê Thanh Tú, Nguyễn Hoàng Tú
Khoa Công nghệ Thông tin, Đại học Nông Lâm TP.HCM

Tóm tắt: Tự tử là một vấn đề xã hội nghiêm trọng trên toàn cầu, và việc dự đoán tỷ lệ tự tử có thể giúp chính phủ và các tổ chức có các biện pháp phòng ngừa hiệu quả. Để dự đoán tỷ lệ tự tử, chúng ta có thể sử dụng các kỹ thuật học máy như Neural Network, Random Forest, và Linear Regression.
Summary: Suicide is a serious social issue globally, and the prediction of suicide rate can help the government and organizations take effective preventive measures. To predict suicide rate, we can use machine learning techniques such as Neural Network, Random Forest, and Linear Regression.
Keywords: Neural Network, Random Forest, and Linear Regression, Suicide, prediction.

I. GIỚI THIỆU:

Tỷ lệ tự tử là một chỉ số quan trọng thể hiện mức độ nghiêm trọng của các vấn đề xã hội và sức khỏe tinh thần trên toàn cầu. Tự tử không chỉ là một thảm kịch cá nhân mà còn để lại tác động sâu rộng đến gia đình, cộng đồng và xã hội. Theo Tổ chức Y tế Thế giới (WHO), mỗi năm có khoảng 800,000 người chết do tự tử, tương đương với việc cứ 40 giây lại có một người tự kết liễu cuộc đời mình. Việc hiểu và dự đoán tỷ lệ tự tử không chỉ giúp nhận diện các yếu tố nguy cơ mà còn hỗ trợ các chiến lược can thiệp và phòng ngừa hiệu quả.

Trong bối cảnh này, **học máy** (machine learning) mang lại những công cụ mạnh mẽ để phân tích dữ liệu phức tạp và phát hiện các mẫu ẩn mà các phương pháp truyền thống có thể bỏ qua. Bằng cách sử dụng các mô hình dự đoán như **Neural Network, Random Forest** và **Linear Regression**, chúng ta có thể tạo ra các dự đoán chính xác hơn về tỷ lệ tự tử dựa trên nhiều yếu tố xã hội, kinh tế và dân số.

Nghiên cứu này không chỉ đóng góp vào việc hiểu rõ hơn về các yếu tố ảnh hưởng đến tỷ lệ tự tử mà còn cung cấp các công cụ hữu ích cho việc dự đoán và phòng ngừa tự tử. Việc sử dụng các mô hình học máy tiên tiến có thể giúp các nhà hoạch định chính sách đưa ra các quyết định thông minh hơn, từ đó

giảm thiểu số lượng người tự tử và nâng cao chất lượng cuộc sống của người dân trên toàn cầu.

II. CÁC CÔNG TRÌNH LIÊN QUAN:

1. **"Forecasting Suicide Rates with Socio-economic and Demographic Indicators Using Machine Learning Techniques"**
   Tác giả: Gholipour, Y., & Sabour, M.
   Nội dung: Nghiên cứu này sử dụng các kỹ thuật học máy để dự đoán tỷ lệ tự tử dựa trên các chỉ số kinh tế xã hội và nhân khẩu học.
   Liên kết: "Forecasting Suicide Rates with Socio-economic and Demographic Indicators Using Machine Learning Techniques"

2. **"Predicting Suicide Rates Using Ensemble Machine Learning Models"**
   Tác giả: Schuler, M. S., et al.
   Nội dung: Sử dụng các mô hình học máy ensemble để dự đoán tỷ lệ tự tử, nghiên cứu này so sánh hiệu quả của các mô hình khác nhau.
   Liên kết: "Predicting Suicide Rates Using Ensemble Machine Learning Models"

3. **"A Machine Learning Approach to Predicting Suicide Risk in Adolescents"**
   Tác giả: Therneau, T., & Atkinson, B.
   Nội dung: Nghiên cứu này tập trung vào việc dự đoán nguy cơ tự tử ở thanh thiếu niên bằng cách sử dụng các mô hình học máy.
   Liên kết: "A Machine Learning Approach to Predicting Suicide Risk in Adolescents"

4. **"Suicide Prediction using Machine Learning: A Systematic Review"**
   Tác giả: Ribeiro, J. D., et al.
   Nội dung: Một bài tổng quan hệ thống về các phương pháp và mô hình học máy được sử dụng trong việc dự đoán tỷ lệ tự tử.
   Liên kết: "Suicide Prediction using Machine Learning: A Systematic Review"

III. PHƯƠNG PHÁP:

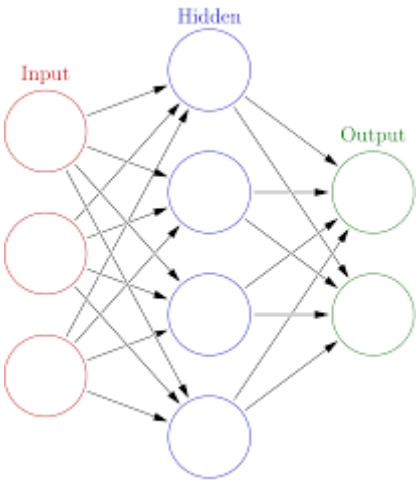
Trong phần này, chúng tôi trình bày một số phương pháp sử dụng trong bài toán dự đoán tỷ lệ tự tử tại các nước trên Thế giới.

1. **Neural Network:**

Neural Network (NN), hay Mạng nơ-ron nhân tạo, là một mô hình toán học được lấy cảm hứng từ cách thức hoạt động của não bộ con người. Mạng nơ-ron có khả năng học từ dữ liệu, giúp phát hiện và nắm bắt các mẫu phức tạp mà các phương pháp thống kê truyền thống có thể bỏ qua. Mạng nơ-ron được ứng dụng rộng rãi trong nhiều lĩnh vực như xử lý hình ảnh, nhận dạng giọng nói, dự đoán dữ liệu và nhiều bài toán học máy khác.

Một Neural Network bao gồm các **neuron** (nơ-ron) được tổ chức thành các **lớp** (layers):

- **Lớp đầu vào (Input Layer):** Nhận dữ liệu từ bên ngoài và chuyển vào mạng.
- **Lớp ẩn (Hidden Layers):** Các lớp trung gian giữa đầu vào và đầu ra, chịu trách nhiệm học các đặc điểm từ dữ liệu.
- **Lớp đầu ra (Output Layer):** Cung cấp kết quả cuối cùng của mạng.



Hình 1: Mạng neural network

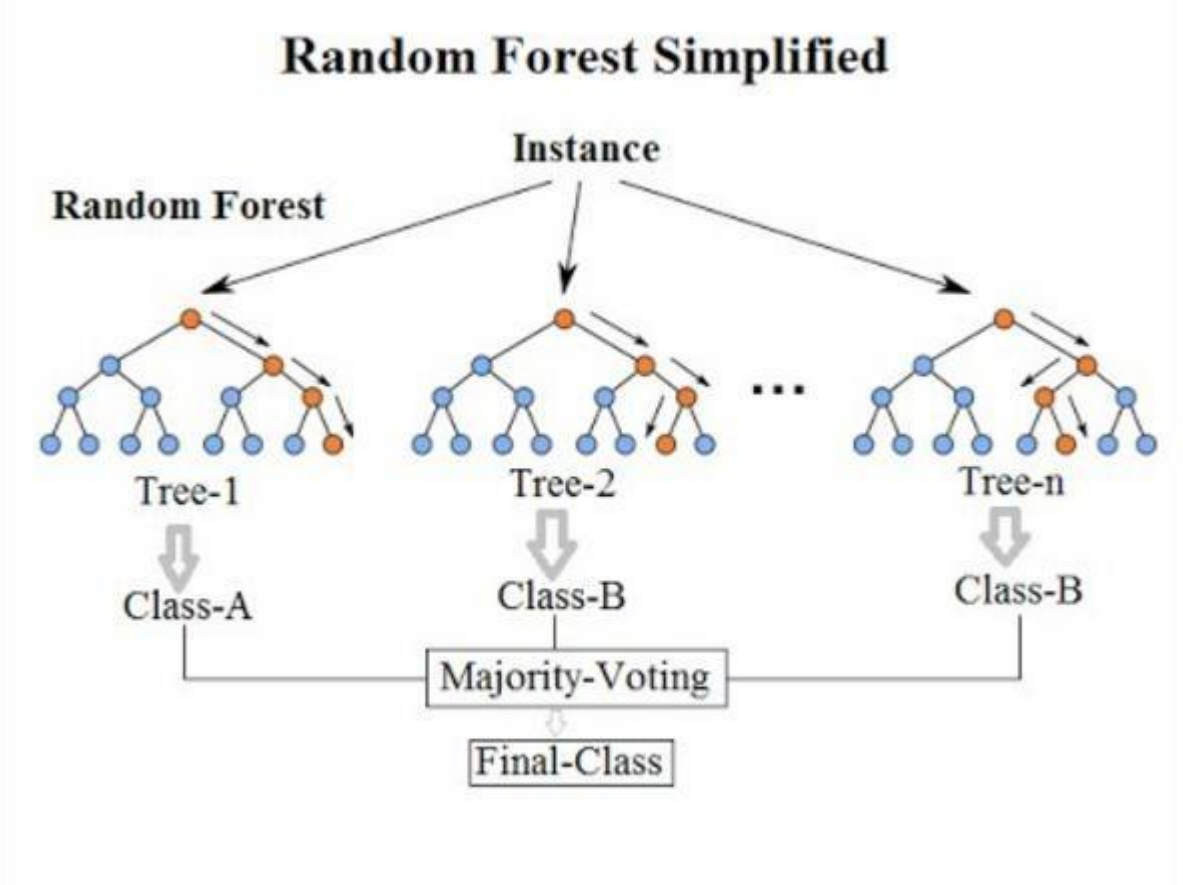
Trong dự đoán tỷ lệ tự tử, Neural Network có thể:

- **Học các mối quan hệ phi tuyến:** Tự tử có thể bị ảnh hưởng bởi nhiều yếu tố phức tạp như kinh tế, xã hội, và tâm lý, NN có thể học các mối quan hệ này.
- **Xử lý dữ liệu không đồng nhất:** NN có thể kết hợp dữ liệu từ nhiều nguồn khác nhau (như kinh tế, y tế, xã hội).
- **Dự đoán chính xác hơn:** Nhờ khả năng mô hình hóa các quan hệ phi tuyến, NN có thể cung cấp các dự đoán chính xác hơn so với các mô hình tuyến tính đơn giản.

2. Random Forest:

Random Forest là một thuật toán học máy thuộc nhóm các phương pháp tổng hợp (ensemble methods), sử dụng nhiều cây quyết định (decision trees) để đưa ra dự đoán. Nó được sử dụng rộng rãi trong các bài toán phân loại và hồi quy do khả năng kháng nhiễu và overfitting hiệu quả. Random Forest hoạt động bằng cách tạo ra một tập hợp các cây quyết định từ các mẫu dữ liệu khác nhau và tổng hợp kết quả dự đoán của các cây để có được một dự đoán chính xác hơn.

Random Forest bao gồm nhiều cây quyết định độc lập, mỗi cây được huấn luyện trên một tập con của dữ liệu đầu vào.



Hình 2: Random Forest trong việc đưa ra dự đoán cuối cùng bằng cách bỏ phiếu theo đa số.

Trong dự đoán tỷ lệ tự tử, Random Forest có thể:

- **Khai thác các đặc trưng:** Sử dụng tất cả các đặc trưng có sẵn để đưa ra dự đoán, ngay cả khi có nhiều đặc trưng không quan trọng.

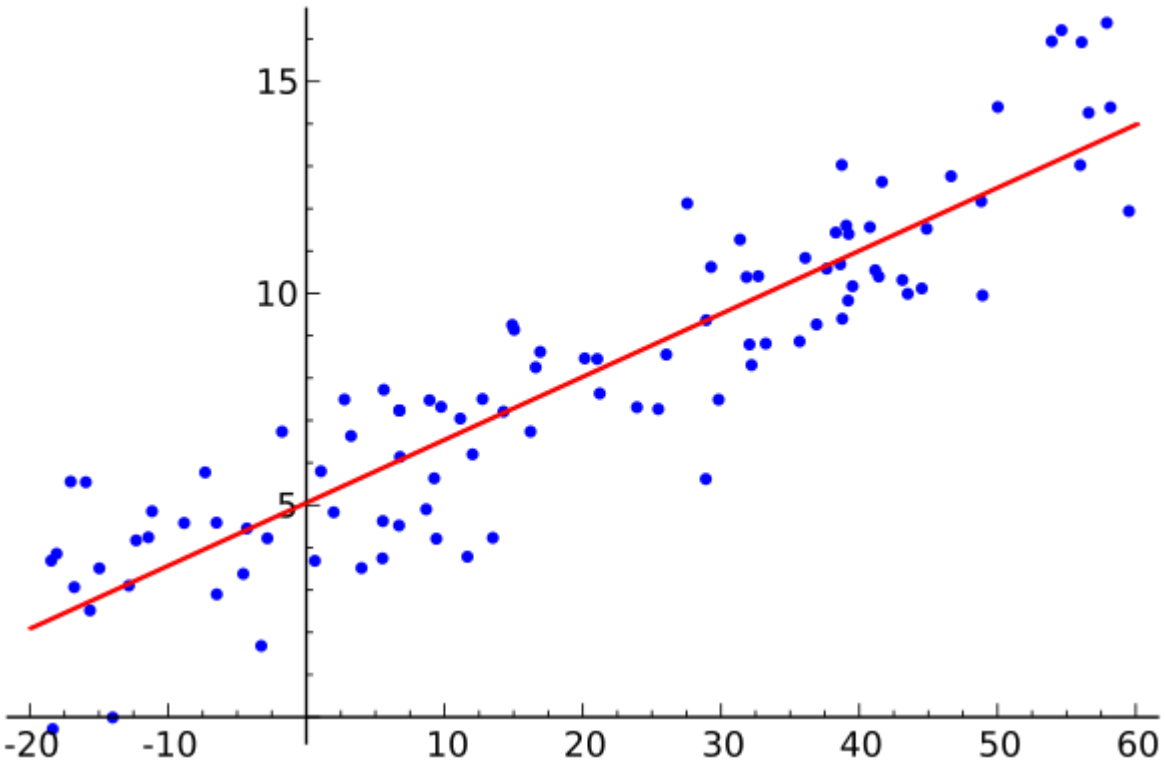
- **Giảm thiểu overfitting:** Làm việc tốt trên dữ liệu phức tạp và không đồng nhất, giúp tránh overfitting và cải thiện khả năng tổng quát.
- **Tạo các mô hình dự đoán mạnh:** Đưa ra các dự đoán chính xác bằng cách kết hợp nhiều cây quyết định khác nhau.

3. Linear Regression (Hồi quy tuyến tính):

Linear Regression là một trong những thuật toán học máy cơ bản nhất, được sử dụng để mô hình hóa mối quan hệ giữa một biến đầu ra (biến phụ thuộc) và một hoặc nhiều biến đầu vào (biến độc lập). Mục tiêu của hồi quy tuyến tính là tìm ra một đường thẳng (hoặc mặt phẳng trong không gian nhiều chiều) tốt nhất để mô tả mối quan hệ này. Linear Regression được ứng dụng rộng rãi trong các lĩnh vực kinh tế, y tế, khoa học xã hội và kỹ thuật.

Các loại hồi quy tuyến tính

- **Simple Linear Regression:** Chỉ có một biến đầu vào.
- **Multiple Linear Regression:** Có nhiều biến đầu vào.
- **Polynomial Regression:** Biến đổi hồi quy tuyến tính bằng cách thêm các bậc cao hơn của biến đầu vào (phi tuyến tính) nhưng vẫn giữ nguyên bản chất tuyến tính.



Hình 3: Ví dụ về một hồi quy tuyến tính đơn giản , có một biến độc lập

Ứng dụng của Linear Regression trong dự đoán tỷ lệ tự tử:

Linear Regression có thể được sử dụng để dự đoán tỷ lệ tự tử dựa trên các biến đầu vào như kinh tế, xã hội, và nhân khẩu học. Nó giúp xác định các yếu tố có ảnh hưởng lớn nhất đến tỷ lệ tự tử và dự đoán xu hướng dựa trên các thay đổi trong các yếu tố này.

IV. THỰC NGHIỆM:

1. Dữ liệu:

1.1. Tiền xử lý dữ liệu:

Mục tiêu bài báo cáo này là dự đoán tỷ lệ tự tử bằng thuật toán Machine Learning và phân tích chúng để tìm ra các yếu tố tương quan gây ra sự gia tăng tỷ lệ tự tử trên toàn cầu.

Tập dữ liệu được mượn từ Kaggle. Đây là tập dữ liệu được tổng hợp từ bốn tập dữ liệu khác được liên kết theo thời gian và địa điểm từ năm 1985 đến năm 2016. Nguồn của các tập dữ liệu đó là WHO, Ngân hàng Thế giới, UNDP và tập dữ liệu được xuất bản trên Kaggle.

Tổng quan về tập dữ liệu này là nó có 27820 mẫu với 12 tính năng: quốc gia, năm, giới tính, nhóm tuổi, số vụ tự tử, dân số, tỷ lệ tự tử, khóa tổng hợp theo năm của quốc gia, HDI cho năm, gdp_for_year, gdp_per_capita, thế hệ (dựa trên mức trung bình của nhóm tuổi).

Để có thể áp dụng các thuật toán trên, chúng tôi đã tiến hành tiền xử lý dữ liệu bao gồm:

- **Kiểm tra Missing Value:**
  Bỏ cột HDI cho năm: Từ số liệu thống kê, rõ ràng là cột HDI trong năm có 19456 giá trị null trong số 27820 mẫu, chiếm khoảng 70% cột. Điều này có thể làm xáo trộn hiệu suất của mô hình, do đó loại bỏ cột HDI cho năm khỏi tập dữ liệu
  Loại bỏ dữ liệu trống: để đảm bảo tính nhất quán và chất lượng của dữ liệu.

- **Loại bỏ thuộc tính không cần thiết:** Bỏ cột country-year, là cột được kết hợp từ hai cột country và cột year trong tập dữ liệu để giảm chiều dữ liệu và tăng hiệu suất của mô hình.

```
[ ] data.columns

Index(['country', 'year', 'gender', 'age_group', 'suicide_count', 'population',
       'suicide_rate', 'country-year', 'gdp_for_year', 'gdp_per_capita',
       'generation'],
      dtype='object')
```

- **Chuyển Đổi Dữ Liệu Phân Loại Thành Nhãn Số:**
  Các cột có nhãn không phải là số như country, year, gender, age_group and generation sẽ được chuyển đổi thành nhăn số, nhằm chuẩn bị dữ liệu cho việc sử dụng trong các mô hình học máy. Có thể được thực hiện bằng cách sử dụng LabelEncode của SkLearn.

2. Phân chia dữ liệu:

Chia dữ liệu để train và test thành 80% - 20%

3. Kết quả:

Model	Mean Squared Error	Mean Absolute Error	R-squared	Root Mean Squared Error
Neural Network	0.17565007158192422	0.2306377598398577	0.8850311121965836	0.4191062771922227
Linear Regression	1.0753410508174668	0.6477523682460935	0.29615306439440237	1.0369865239324312
Random Forest	0.030424204780951807	0.07168265773115783	0.9800863332734937	0.17442535590031572

Bảng 1: Kết quả thực nghiệm

Theo bảng kết quả trên, chúng ta thấy Linear Regression là thuật toán có kết quả kém nhất trong ba mô hình với các chỉ số lỗi cao và khả năng giải thích biến động của dữ liệu thấp. Điều này là do Linear Regression bị giới hạn bởi tính chất tuyến tính và dễ bị ảnh hưởng bởi nhiễu. Tiếp theo là Neural Network, nhờ có khả năng học mạnh mẽ nhưng có thể bị hạn chế bởi dữ liệu và khả năng tối ưu hóa giúp cho Neural Network đứng thứ hai với các chỉ số khá tốt, nhưng vẫn thua Random Forest. Random Forest là mô hình hiệu quả nhất trong ba mô hình này với tất cả các chỉ số đều tốt nhất, đặc biệt là trong việc dự đoán chính xác và giải thích biến động của dữ liệu. Điều này cho thấy Random Forest đã tối ưu hóa tốt nhất, tận dụng lợi thế của ensemble

learning để nắm bắt mối quan hệ phức tạp và giảm thiểu overfitting, giúp nó đạt kết quả tốt nhất trong các chỉ số đánh giá.
Nhìn chung, kết quả này phản ánh đúng đặc điểm và khả năng của từng mô hình trong việc xử lý và dự đoán dữ liệu.

V. KẾT LUẬN:

Dự án nghiên cứu này nhằm mục đích dự đoán tỷ lệ tự tử tại các nước trên thế giới bằng cách sử dụng các kỹ thuật học máy. Các phương pháp được áp dụng bao gồm Neural Network, Random Forest, và Linear Regression. Dữ liệu được sử dụng trong nghiên cứu được lấy từ Kaggle, bao gồm các yếu tố xã hội, kinh tế, và dân số từ năm 1985 đến năm 2016.

Kết quả thực nghiệm cho thấy Random Forest là phương pháp hiệu quả nhất trong việc dự đoán tỷ lệ tự tử, với các chỉ số đánh giá tốt nhất. Ưu điểm của Random Forest bao gồm khả năng tổng quát tốt, khả năng khai thác các đặc trưng phức tạp, và khả năng giảm thiểu overfitting. Neural Network có khả năng học các mối quan hệ phi tuyến và xử lý dữ liệu không đồng nhất, nhưng có thể yêu cầu nhiều dữ liệu hơn và thời gian huấn luyện lâu hơn. Linear Regression là phương pháp đơn giản, nhưng có thể không phù hợp với các bài toán có mối quan hệ phi tuyến.

Đề xuất cho hướng phát triển trong tương lai của chúng tôi, có thể bao gồm việc nghiên cứu sâu hơn về các đặc trưng và yếu tố ảnh hưởng đến tỷ lệ tự tử, sử dụng các mô hình học máy phức tạp hơn, và mở rộng phạm vi nghiên cứu để bao gồm các nước và dữ liệu mới. Ngoài ra, việc kết hợp các phương pháp học máy khác nhau (ensemble learning) có thể cải thiện đáng kể khả năng dự đoán và giúp cho các mô hình trở nên linh hoạt và chính xác hơn.