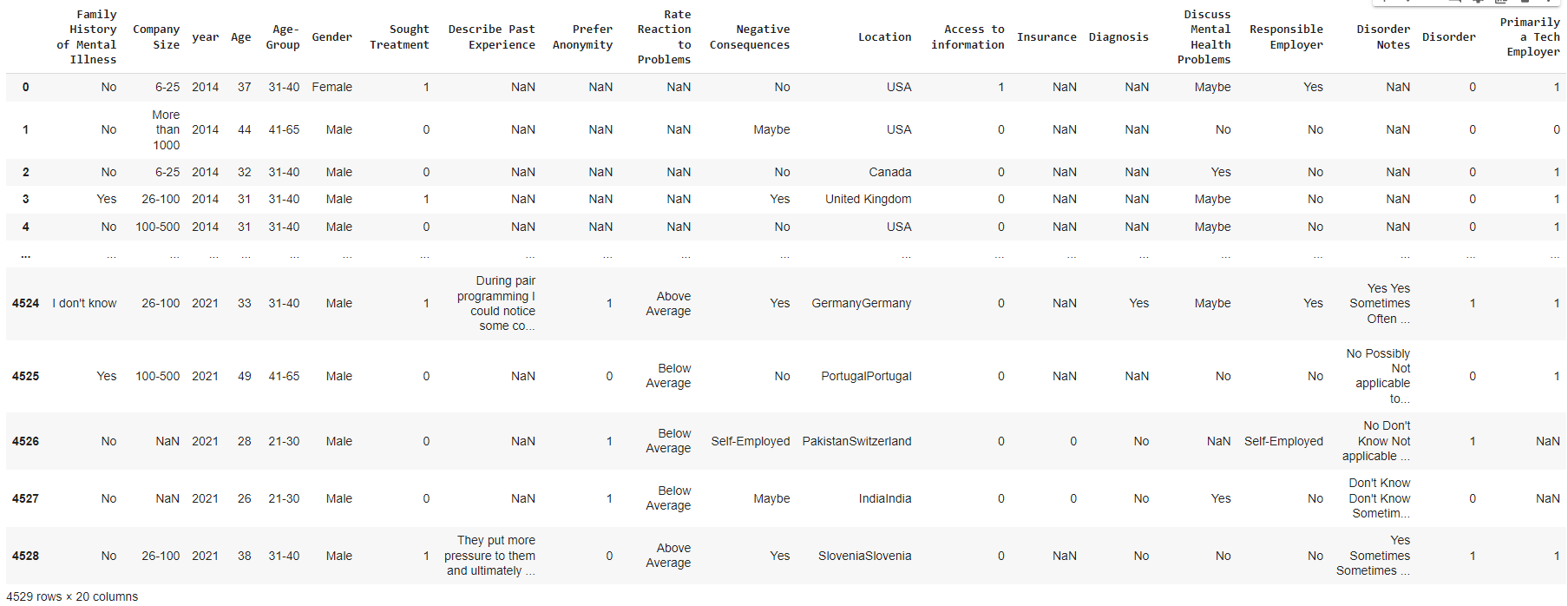
Đồ án được chia làm ba phần chính:

* Data cleaning
* Data analysis
* MachineLearning

Tệp dữ liệu được lấy ở Open Sourcing Mental Illness, bao gồm năm 2014,2016,2017,2018,2019,2020,2021 (không có dữ liệu năm 2015)

Dữ liệu 2020, 2021 quá bé chỉ hơn 100 row, chạy thử cho thấy độ chính xác giảm tại các mô hình

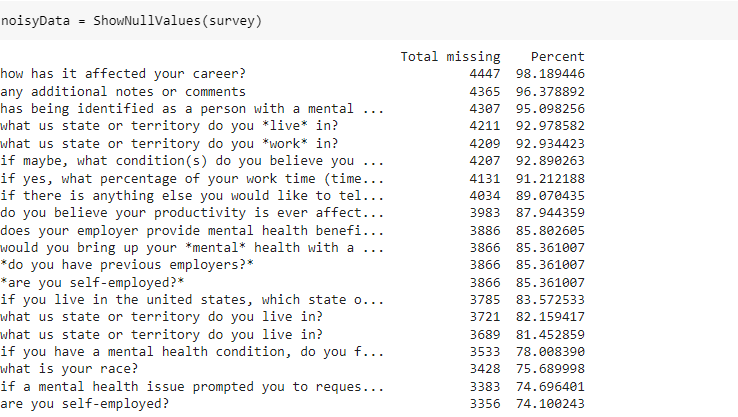
Chỉ dùng 2014,2016,2017,2018,2019



Dữ liệu khi được gộp các năm với nhau.

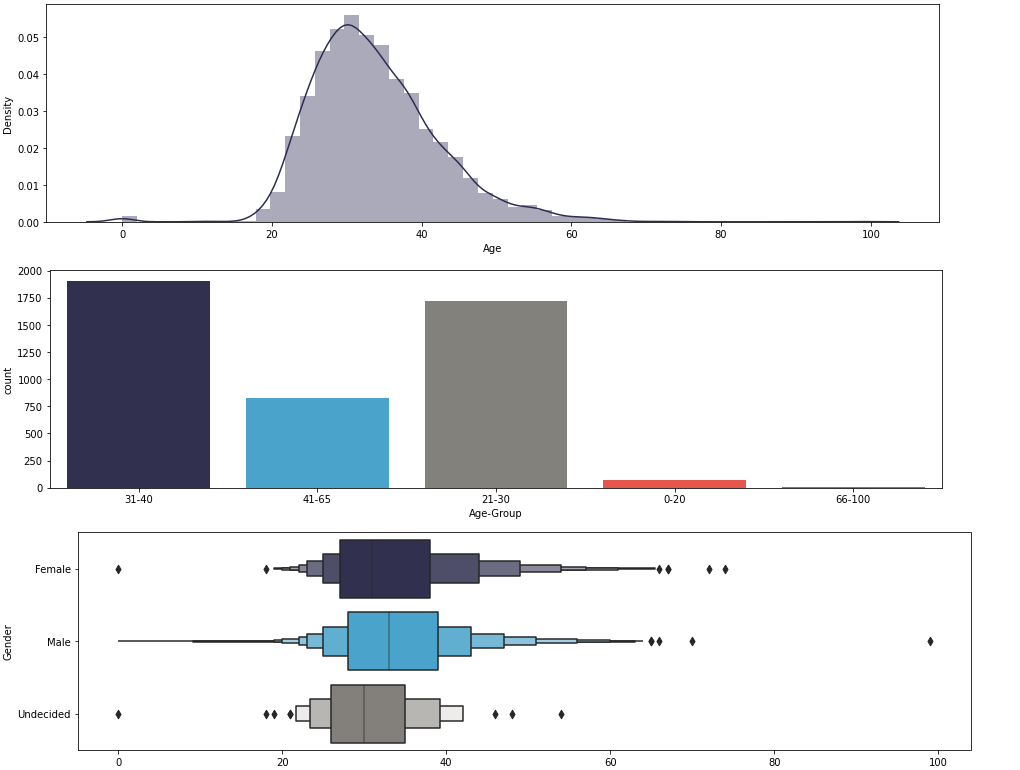
Dữ liệu sau khi thu được cần được làm sạch bằng các cách:

* Đặt lại tên cột
* Bỏ các cột không cần thiết
* Chuẩn hóa dữ liệu về dạng phù hợp
  + Tuổi
    - Loại bỏ các trường hợp <10 và >100
    - Chia độ tuổi thành các nhóm ("0-20", "21-30", "31-40", "41-65", "66-100")
  + Giới tính: Dữ liệu giới tính chứa nhiều kiểu và font chữ khác nhau cần chuyển về một dạng
  + Sought treament: chuyển về dạng boolean
  + Anonymous: chuyển về dạng boolean
  + …



Dữ liệu còn lại sau khi được làm sạch

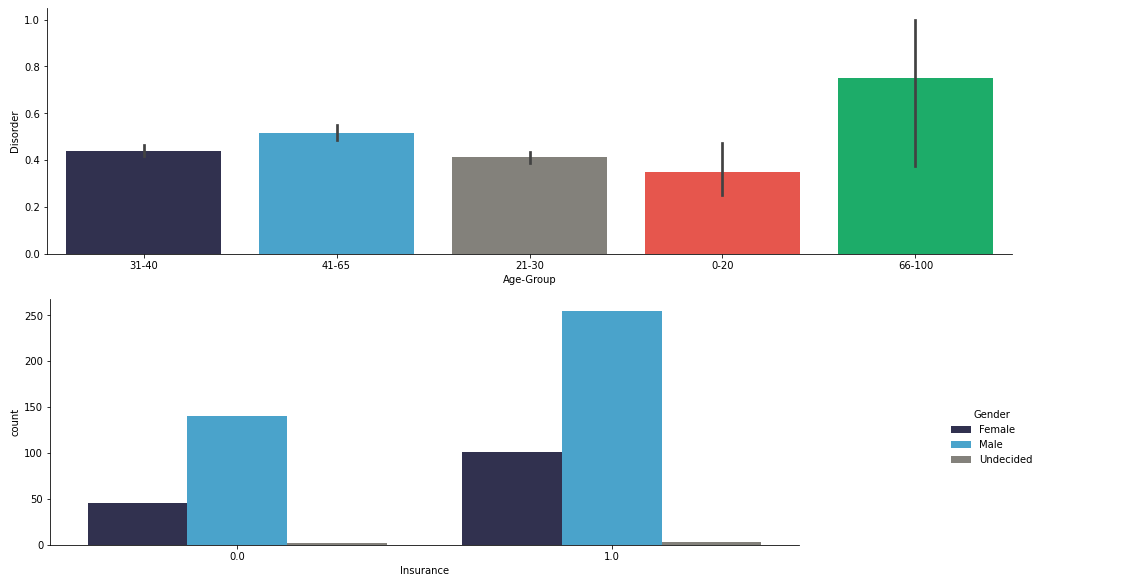
II. Phân tích dữ liệu.

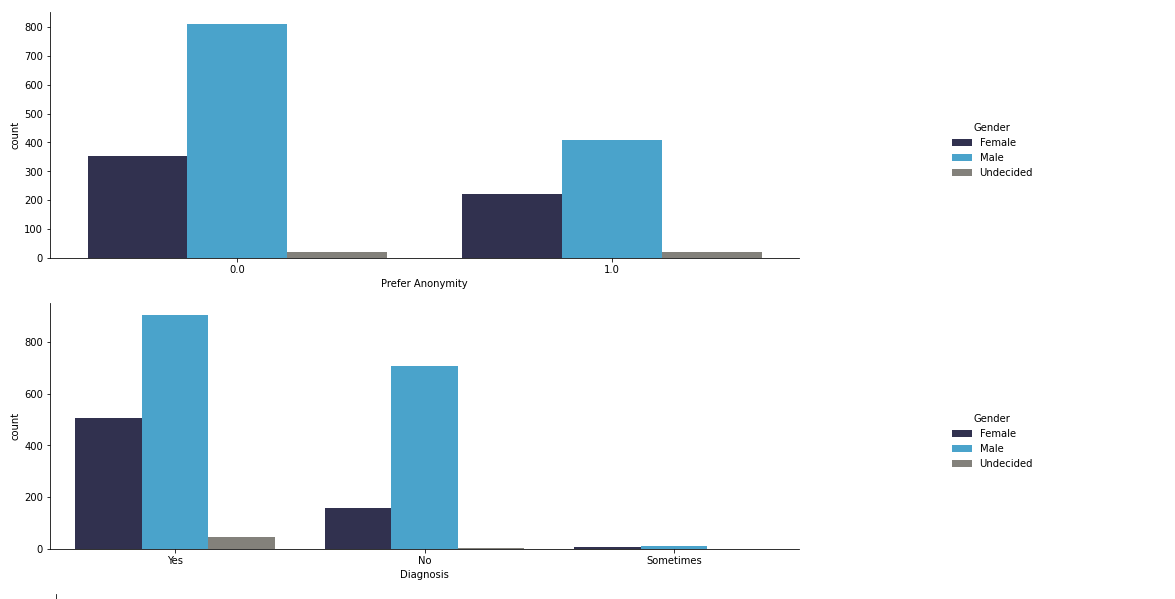


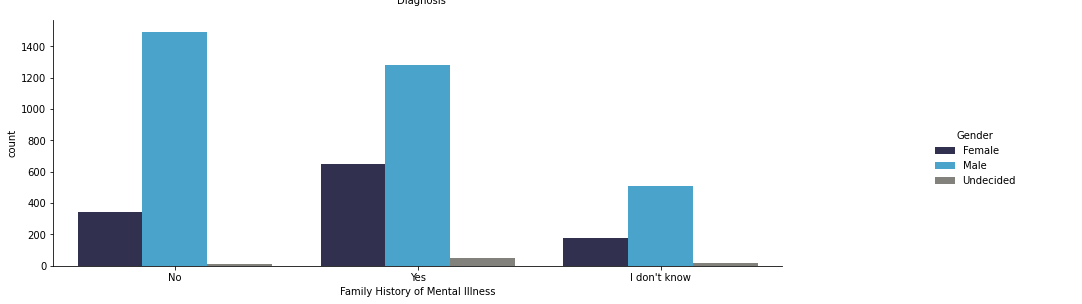
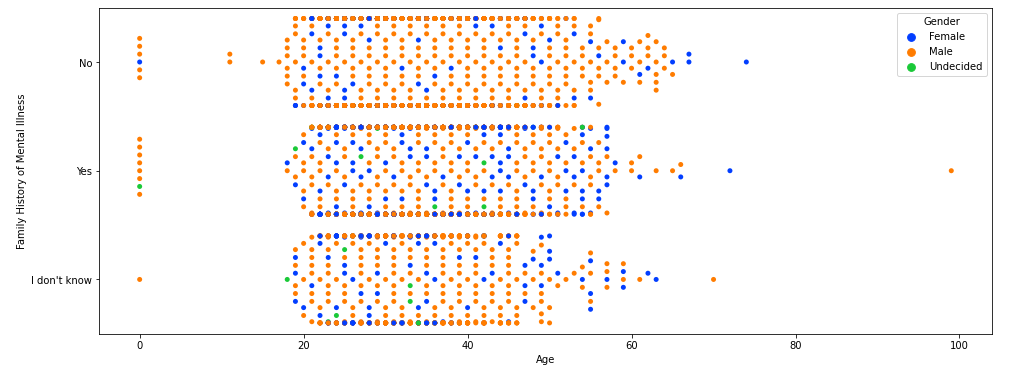
Phân bổ theo độ tuổi và giới tính.

Nhận xét tổng quan, tệp mẫu tập trung ở độ tuổi từ 20-40

Phân bổ theo độ tuổi và giới tính ở các biến Disorder, Insurance, Prefer Anonymity, Diagnosis, Family History of Mental Illness.



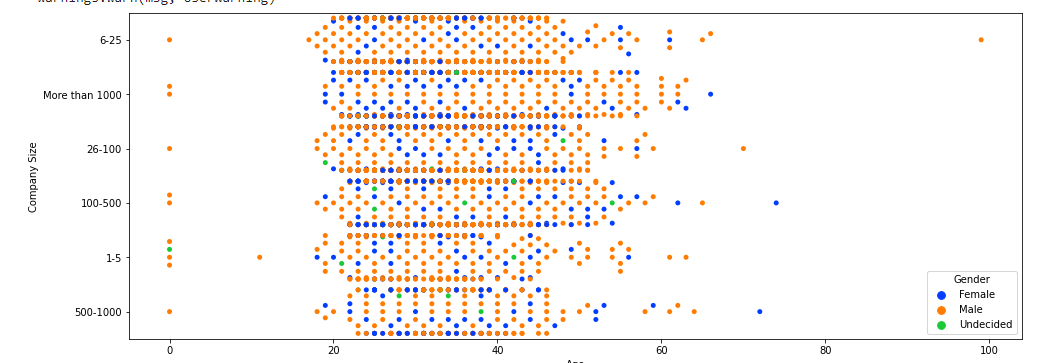


Một số xu hướng của dữ liệu

* Hầu hết mẫu đều đến từ USA hoặc các nước đang phát triển nên đều có bảo hiểu sức khỏe tâm thần
* Ẩn danh là yêu tố được mong muốn, tuy nhiên nó không phải là yếu tố ảnh hưởng đến sự trợ giúp.
* Phần lớn người được hỏi cho biết bản thân mắc rối loạn về tâm thần như PTSD,...
* NHiều phụ nữ cho biết yếu tố di truyền liên quan đến bệnh tâm thần
* Những người không xác định giới tính có tỷ lệ mắc bệnh tâm thần cao hơn nhiều

Kích cỡ của công ty có ảnh hưởng tới sức khỏe tâm thần



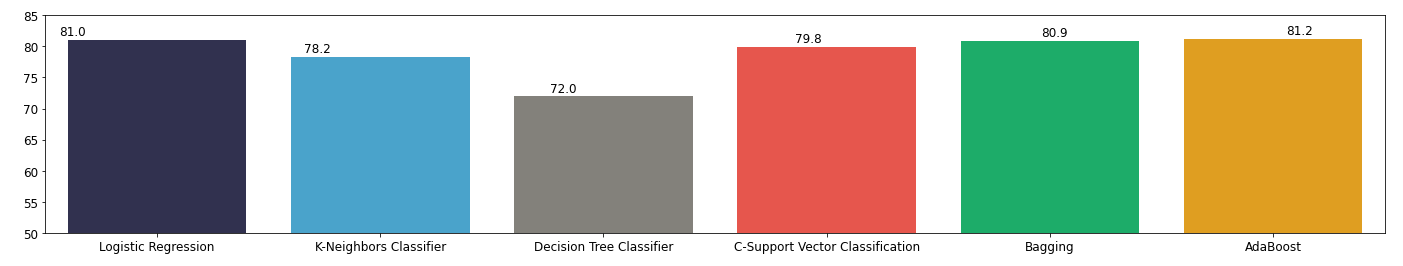
Hầu hết những người được hỏi hoặc người lao động dường như không hài lòng với cách điều trị các vấn đề sức khỏe tâm thần tại công ty hoặc nơi làm việc của họ.

* Ở tất cả các quy mô công ty, những người được hỏi cảm thấy họ có thể được trợ giúp nhiều hơn.
* Không biết được đưa ra một vấn đề sức khỏe có phải là một ý kiến hay.
* Ở công ty lớn, đưa ra vấn để sức khỏe tâm thần có thể mang lại hâuk quả tiêu cực.
* Công ty nhỏ cung cấp ít nguồn lực hỗ trợ về sức khỏe tâm thần
* Nhưng công ty nhỏ giải quyết vấn đề về sức khỏe tâm thần tốt hơn

Từ dữ liệu được phân tích ở trên lựa chọn các biến đưa vào mô hình training bao gồm: { 'Family History of Mental Illness', 'Company Size', 'year', 'Age', 'Age-Group', 'Gender', 'Prefer Anonymity', 'Rate Reaction to Problems', 'Negative Consequences', 'Access to information', 'Diagnosis', 'Insurance', 'Discuss Mental Health Problems', 'Responsible Employer', 'Disorder', 'Primarily a Tech Employer'}.

Chia dữ liệu: 60% dành cho training và 40% testing

Ta sẽ đào tạo trên các mô hình : Logistic Regression, K-Neighbors Classifier, Decision Tree Classifier, C-Support Vector Classification, Bagging meta-estimator, AdaBoost



Độ chính xác của từng mô hình

Mô hình có độ chính xác tốt nhất là Ada Boot 81.2%