TITANIC - MACHINE LEARNING

GIỚI THIỆU VÀ MỤC TIÊU DỰ ÁN

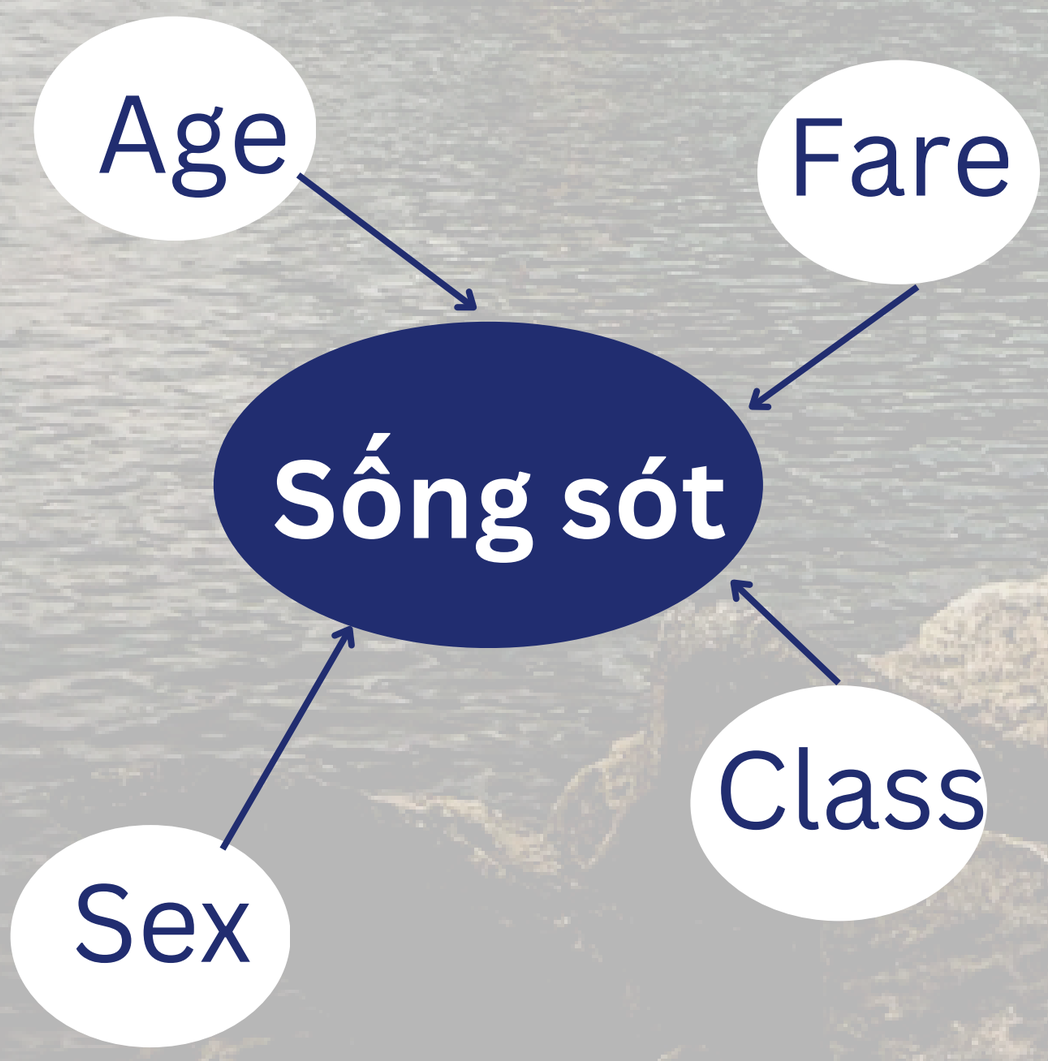
- MỤC TIÊU: DỰ ĐOÁN KHẢ NĂNG SỐNG SÓT TRÊN DỮ LIỆU CỦA KAGGLE
- DỮ LIỆU: TRAIN(~891), TEST(~418)

Thách thức chính:

xử lý tiền dữ liệu và nhiều dữ liệu bị thiếu

Phân tích khám phá dữ liệu

- Giới tính
- Hạng vé
- Độ tuổi
- Gia đình
- Cảng tàu
- Tên riêng



Tiền xử lý và kỹ thuật đặc trưng

Xử lí dữ liệu thiếu

Age : Mean/Median

Cabin: bỏ cột (drop)

Embased: S (phổ biến)

Tạo dạng đặc trưng mới:

Familysize: SibSp + Parch +1

Title: trích ra từ name

Chuẩn hóa:

Title: sử dụng từ name

Age & Fare: Scaling

Tên modle	Xác suất
KNeighborsClassifier	82,94
DecisionTreeClassifier	80,36
RandomForestClassifier	80,92
GaussianNB	80,13
SVC	83,39